Entwicklung von rangbasierten Kriterien und Methoden zur Optimierung der Normalisierung von Genexpressionsexperimenten am Beispiel membranbasierter cDNA-Arrays

Dissertation
Zur Erlangung des akademischen Grades
doctor rerum naturalium (Dr.rer.nat.)

vorgelegt dem Rat der Pharmazeutisch-biologischen Fakultät
der Friedrich - Schiller - Universität

von
Dipl.Chem. Torsten C. Kroll
geboren am 30. September 1971 in Schkeuditz

. . . gewidmet all' denjenigen, die der Fertigstellung dieser Arbeit entgegengesehnt haben, allen voran mein Betreuer Stefan und meine Frau Theresa . . .

1.Gutachter: _____

2.Gutachter: _____

3.Gutachter: _____

Tag der Verteidigung: _____

Inhaltsverzeichnis

Abkürzungsverzeichnis

Im Text

^{32}P	Phosphor 32 - radioaktives Isotop
^{33}P	Phosphor 33 - radioaktives Isotop
AGM	Asymmetrisch Gestutztes Mittel
Anh.	Anhang
ATCC	American Type Culture Collection
BMP2	Bone Morphogenic Protein 2
cDNA	complementary DNA (Komplementäre DNA)
Cy3	Fluoreszenzfarbstoff Cy3 von Molecular Probes Inc.
Cy5	Fluoreszenzfarbstoff Cy5 von Molecular Probes Inc.
DNA	desoxy ribonucleic acid (Desoxyribonukleinsäure)
GBD	global background dots - Hintergrundmethode[M&M]
GC-Gehalt	Gehalt einer Sequenz an Guanin/Cytidin-Basen
GE	Genexpression
GEA	Genexpressionsanalyse
GEDA	Genexpressionsdatenanalyse
GIR	blobal image region - Hintergrundmethode[M&M]
Gl.	Gleichung
GM	Gestutztes Mittel
GRISP	Gleichrangintensitätsscatterplot
HAA 1.2	Human atlas array (Clontech)
Hyb.	Hybridisierung
k.A.	keine Angabe
komp.	kompetetiv
LDR	local dot ring - Hintergrundmethode[M&M]
LGR	local grid ring - Hintergrundmethode[M&M]
linSP	linearer Scatterplot
logSP	logarithmischer Scatterplot
M&M	Material und Methoden
MA-Plot	Microarray-Plot
MNS	mode of non spot - Hintergrundmethode[M&M]
mRNA	Boten-RNS (messenger RNA)
MW	Mittelwert
n.a.	nicht anwendbar
NA	nucleic acids (Nukleinsäuren)
NN	Nearest Neighbor (Nächste Nachbarn)
OGM	Oberes Gestutztes Mittel
Oligo	DNA-Oligomer
ORF	open reading frame (Offener Leserahmen)
PCR	polymerase chain reaction (Polymerase Kettenreaktion)
polyA-RNA	RNA, welche am 3' Ende eine Polyadenosinsequenz enthält
RID	Rang - Intensitäts - Diagramm
RIK	Rang - Intensitäts - Kurve
RISA	Rang-Intensitäts-Standardabweichung
RNA	ribonucleic acid (Ribonukleinsäure)
rRISA	relative Rang-Intensitäts-Standardabweichung
rRNA	ribosomale RNA
RT-PCR	Reverse Transkription mit anschließender PCR
s.o.	siehe oben
sGM	Symmetrisch Gestutztes Mittel
SNP	single nucleotide polymorphism (Einzelbasenpolymorphismus)
SSM	satellite spot method - Hintergrundmethode[M&M]
TF	Transkriptionsfaktor
tRNA	transfer RNA (Transfer- RNA)
UGM	Unteres Gestutztes Mittel
UV	Ultraviolettes Strahlung
WBD	global background dots - Hintergrundmethode[M&M]
WIR	weighted image regions - Hintergrundmethode[M&M]
z.B.	zum Beispiel

Symbole und Indizes in Formeln

ΔS_{sym}	Symmetrieterm
$\Delta_R G$	Freie Energie
$\Delta_R H$	Enthalpie
$\Delta_R S$	Entropie
$\nu_{label,hyb}$	Gesamtanzahl signalgebender Markierungen der RNA/cDNA- Moleküle der zu bestimmenden Spezies welche an die Sonden des Arrays hybridisiert haben

ν_{label}	Gesamtanzahl signalgebender Markierungen der RNA/cDNA -Moleküle der zu bestimmenden Spezies
A	Aufreinigung
A	in chemischen Formeln: NA in der Probe mit einer bestimmten Sequenz
A_{rc}	in chemischen Formeln: NA mit einer reverse komplentären Sequenz zu A
B	in chemischen Formeln: NA in der Probe mit einer bestimmten Sequenz
c	concentration (Konzentration)
$card$	Kardinalität (Anzahl der Elemente einer Menge)
$\Delta\gamma$	Fehler des Biasterms
$\Delta\kappa$	Fehler des Normalisierungsquotient
Δs_n	Fehler des normalisierten Signals
Δs_p	Fehler des Einzelwertes
Δs_r	Rangwertfehler
erf	Errorfunktion
exp	Exposition
f	Einflußfaktor
f_{amp}	Signalverstärkungsfaktor
f_{einbau}	Einbaurate
f_q	Quantenausbeute
$frac$	Restfunktion
γ	Biasterm
H	Arrayhybridisierung
h	Skalierungsfaktor
H_{init}	Initiationsterm
I	Intensität
int	Integerfunktion
I_r	Rangintensität
k	Geschwindigkeitskonstante
K	Gleichgewichtskonstante
κ	Normalisierungsquotient
L	Markierung
$label$	label (Markierung)
M	Markierungsmessung
mm	mismatch (unvollständige Sequenzkomplementarität)
N	Anzahl (z.B. Moleküle)
n	Stoffmenge
N_A	Avogadro- Konstante
n_{A_0}	Ursprungsmenge an Probe A
N_{Abbau}	Anzahl in den Abbauprozess überführter Moleküle
N_{Aufbau}	Anzahl produzierter Moleküle
n_{B_0}	Ursprungsmenge an Probe B
n_{label}	Anzahl reaktiven Markierungsmoleküle
n_{mRNA}	Menge an zu messender RNA Stoffmenge, Molekülanzahl
n_{pur}	durch die Aufreinigung veränderte RNA-Menge
n_{S_0}	Ursprungsmenge an Sonde S
$n_{sampled}$	durch die Probennahme veränderte RNA-Menge
o	oberere Stutzung
P	Probennahme
pm	perfect match (vollständige Sequenzkomplementarität)
Q	Bildquantifizerung
$quantil$	Quantilfunktion
R	molare Gaskonstante
$rfun$	Rangfunktion
rgb	Rangbindungsfunktion
rgw	Rangwertfunktion
$rund$	Rundungsfunktion
S	in chemischen Formeln: Sonden-NA mit einer bestimmten Sequenz
S	Signal
σ_r	Rangwertfehler
s_p	Signal
s_n	normalisiertes Signal
s_r	ranggeordnetes Signal
S_{init}	Initiationsterm
s_{label}	Gesamtsignal der Markierungen, welches durch den Flächendetektor gemessen wurde
s_{quant}	Gesamtsignal des Bereiches auf dem Array (respektive Bilddatei) welcher der Probenkomplentären Sonde zugeordnet wird, abzüglich diverser Störsignale und Hintergrundbias
T	Temperatur
t	time (Zeit)
T_m	melting temperature (Schmelztemperatur)
tot	total
u	untere Stutzung
V	Volumen
V_{probe}	Probenvolumen
V_{sonde}	Sondenvolumen
V_{total}	Gesamtvolumen

Kapitel 1

Einleitung

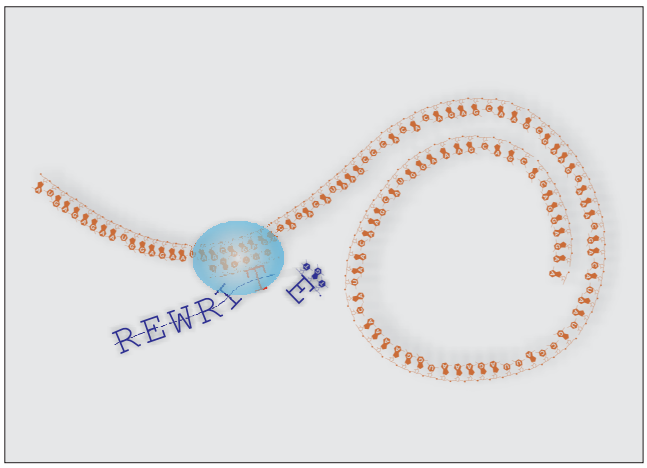


Abbildung 1.1: „Eine simple Botschaft“

Nachdem das menschliche Genom und die einiger anderer wichtiger Spezies fast vollständig sequenziell kartiert sind [IHGSC2001], [Venter2001], bleibt die Aufgabe übrig, diese Karten zu annotieren und mit biologischer Bedeutung zu füllen [Stein2001]. Die vollständige funktionelle Analyse der Sequenzinformation ist damit die nächste große Herausforderung. Die schiere Menge der erzeugten Information läßt jedoch traditionelle Methoden der funktionellen Charakterisierung aller gefunden Gene zur Sisyphos-Arbeit ausarten. Andererseits ermöglicht dieses neue Wissen auch, herkömmliche Methoden zu verbessern und zu automatisieren, und somit neue Impulse zu setzen. Eine der Techniken, die dadurch in den letzten 10 Jahren entwickelt werden konnte, ist die Genexpressionsanalyse mittels cDNA-Arrays. Durch diese ist es möglich den Transkriptionsstatus von Zellen zu bestimmen, das heißt die molekularen Mengenverhältnisse tausender verschiedener mRNA-Spezies gleichzeitig zu messen. Diese Methode erlaubte erstmals die kostengünstige genomweite Analyse der Genexpression verschiedener Zellzustände und Zelltypen [Schena1995], [Velculescu1995]. Neben diesen ersten herausragenden Erfolgen der Methode zeichneten sich aber auch verschiedene technologische Herausforderungen ab. Zum einen ist die eigentliche Messung für den Experimentator relativ einfach auszuführen, verlangt aber einen großen Vorbereitungs- und Auswerteaufwand. Zum anderen sind die verwendeten Techniken sehr komplex und dadurch fehleranfällig. Weiterhin werden immense Datenmengen generiert, die nicht mehr manuell verarbeitet, kontrolliert und

verifiziert werden können. Doch gerade die Lösung dieser neuartigen Probleme ist entscheidend für die biologische Signifikanz der publizierten Interpretationen. Die vorliegende Arbeit beschäftigt sich daher mit den folgenden grundlegenden Fragestellungen.

- Welche Parameter beeinflußen die gewonnenen Expressionsdaten und in welcher Weise? ⇒ Systemanalyse
- Welche dieser Parameter sind kritisch und führen zu falschen Daten? ⇒ Fehlerbetrachtung
- Können einfache Methoden zur Vergleichbarmachung (Normalisierung) gefunden werden? ⇒ Normalisierungsmethoden

1.1 Biologische Grundlagen - Genexpression und Genregulation

Der menschliche Körper, als für uns wichtigster multizellulärer Organismus, besteht aus mehr als 1 Billion (10^{12}) Zellen. Über 100 Zelltypen sind zu einem hochkomplexen System verbunden. Fast jede dieser Zellen enthält genau den gleichen Bauplan in Form von Desoxyribonukleinsäurepolymeren (DNA), die wiederum in Form von Chromosomen vorliegen. Dadurch sind die Zellen eines Organismus weitgehend genetisch identisch. Und doch entwickeln sich viele Zellen extrem unterschiedlich. Angefangen mit der omnipotenten befruchteten Keimzelle, entwickeln sie sich zu hochspezialisierten Zellen mit den verschiedensten Funktionen. Jede Spezialisierung ist eine der möglichen Interpretationen der Erbinformation. Den Mechanismus, mit der aus dem gleichen Standardbauplan unterschiedliche Zellen entwickelt werden, nennt man Differenzierung. Jede einzelne Zelle, auch in einzelligen Lebewesen, enthält verschiedene Mechanismen zur Regulation der Information, welches Gen auf der DNA in Protein übersetzt wird und wie viel Kopien dieses Proteins gebildet werden. Das Verständnis dieser Prozesse ist demzufolge eine notwendige Voraussetzung, um z.B. krankhafte Fehlinterpretationen dieses Bauplanes zu erkennen und möglicherweise gezielt zu beeinflussen.

Um diese Veränderungen zu beschreiben, benötigt man eine Definition des Differenzierungszustandes einer Zelle. Dieser Zustand wird charakterisiert durch den mikroskopischen Phänotyp und den Molekularzustand der Zelle. Auf mikroskopischer Ebene sind Zellform, -aufbau, -komponenten und Funktion innerhalb des gesamten Gewebes/Organismus wichtige Beschreibungsgrößen. Auf molekularer Ebene ist es die Gesamtheit aller Moleküle, die eine Zelle bilden. Da die lebende Zelle kein statisches System ist, gibt es zeitliche Veränderungen der molekularen Zusammensetzung. In diesem Sinne werden sich alle Zellen unterscheiden, auch wenn sie mikroskopisch denselben Aufbau haben. Die DNA-Sequenz (Genom) ist, wie schon beschrieben, im multizellularen Organismus hoch konservativ. Sie scheidet damit zur ausschließlichen Beschreibung des molekularen Phänotypes aus. Der mikroskopische Phänotyp ist immer strukturell und funktionell differenziert. Charakteristisch müssen daher struktur- und funktionsprägende Moleküle sein. Das sind im wesentlichen Proteine. Sie sind außerdem die Endprodukte der Genexpression. Demzufolge ist die Gesamtheit aller Proteine (Proteom) einer Zelle charakteristisch für ihren molekularen Differenzierungszustand.

Prokaryoten Jede Differenzierung einer Zelle bedeutet das Proteom der Zelle zu verändern. Die Regulation dieser Veränderung findet auf allen Ebenen statt, die zwischen der genetischen Information und dem letztendlich produzierten Protein liegen. In prokaryotischen Organismen ist dieser Prozeß relativ einfach organisiert (Abbildung 1.1). Alle Regulationsebenen liegen im Zytosol. Die (Gen-)Produkthemmung kann direkt erfolgen, da jedes potentiell interagierende Protein Zugang zur DNA, zur Transkriptase bzw. zu den Transkriptionsfaktoren (Aktivatoren sowie Repressoren) hat.

Eukaryoten In eukaryotischen Organismen ist dieser Vorgang sehr viel komplexer (Abbildung 1.1). Die Erbinformation liegt im Zellkern als Chromatin vor. Das Chromatin stellt einen DNA-Protein Komplex

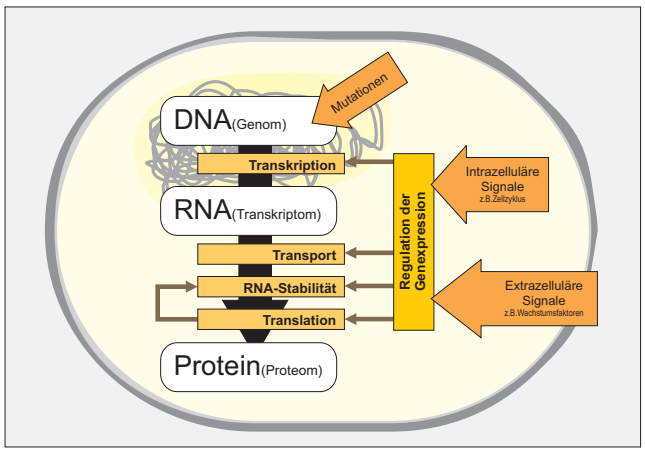


Abbildung 1.2: Regulation der Genexpression bei Prokaryoten

dar, auf den die jeweilige chromosomale DNA auf Histonoktamere aufgewunden vorliegt. Die Struktur des Chromatins selbst und Proteine, die auf diese Struktur Einfluß nehmen können, regeln den sterischen Zugang zu den Genen. Gene bestehen aus mehreren kodierenden Sequenzeinheiten (Exons), die durch eingefügte nichtkodierende, aber teilweise regulierende Sequenzen (Introns) unterbrochen werden. Die Regulation auf Transkriptionsebene erfolgt durch Transkriptionsfaktoren (TF). Die allgemeinen TFs legen die Initiationsstelle der Transkription fest. Die ubiquitären „*upstream*"-Faktoren binden an verschiedene stromaufwärtsgelegene regulatorische Sequenzen (Bindungsstellen für TFs). Weiterhin existieren noch induzierbare TFs. Diese sind ähnlich der „*upstream*"-Faktoren, sind jedoch nicht ubiquitär und unterscheiden sich durch ihre regulatorische Funktion. Erst die gemeinsame Bindung aller notwendigen TFs für das jeweilige Gen ermöglicht die Bindung der RNA-Polymerase an die DNA und die Initiation der Transkription.

Doch auch die Information auf DNA-Ebene ist nicht vollständig unverändert, so kann z.B. durch gezielte Methylierungen Veränderungen im Bindungsverhalten des methylierten Sequenzabschnittes erzeugt und die Bindung eines TF verhindert werden.

Ein weiterer Unterschied zu Prokaryoten liegt in der Prozession (Weiterverarbeitung) der mRNA. Das 5'Ende der mRNA wird mit einen Methylguanosin-CAP markiert. An das 3'Ende synthetisiert die Poly-A-Polymerase weitere Riboadenosine zu einem poly-A-Schwanz. Gleichzeitig werden die transkribierten Introns ausgeschnitten. Diesen Prozeß nennt man Spleißen (*splicing*). Dabei können auch Introns ausgeschnitten werden in denen eigentlich exprimierbare Sequenzen liegen. Dadurch entstehen verschiedene Spleißvarianten eines Genes mit unterschiedlicher Funktionalität. Die fertig prozessierte mRNA wird danach aktiv aus dem Zellkern transportiert. Im Zytosol werden die exportierten mRNAs durch Ribosomen gebunden. Durch diese Bindung wird die Translation der mRNA zum Protein gestartet. Danach wird die mRNA je nach Stabilität im Zytosol abgebaut oder wieder translatiert. Das Gen ist damit als Protein fertig exprimiert. Es kann noch posttranslational prozessiert werden (z.B. Abtrennung eines Transportsignalpeptids am Ziel-/Wirkungsort, Glykolisierung).

Alle Schritte der Expression eines Proteins werden durch Zellprozesse reguliert. Die häufigste Regulation der Expression ist die Regulation der Transkription durch Transkriptionsfaktoren. Im Gegensatz zu Prokaryoten sind für den erfolgreichen Transkriptionsstart eines Genes mehrere Transkriptionsfaktoren (TF) verantwortlich. Erst die Kombination mehrerer TFs und der RNA-Polymerase zum funktionsfähigen Transkriptionskomplex ermöglicht die Expression eines speziellen Genes. Aufgrund dieser Kombination ist es möglich, daß relativ wenige TFs viele unterschiedliche Gene regulieren können.

Wie kann nun der molekulare Differenzierungsstatus einer Zelle experimentell bestimmt werden? Die

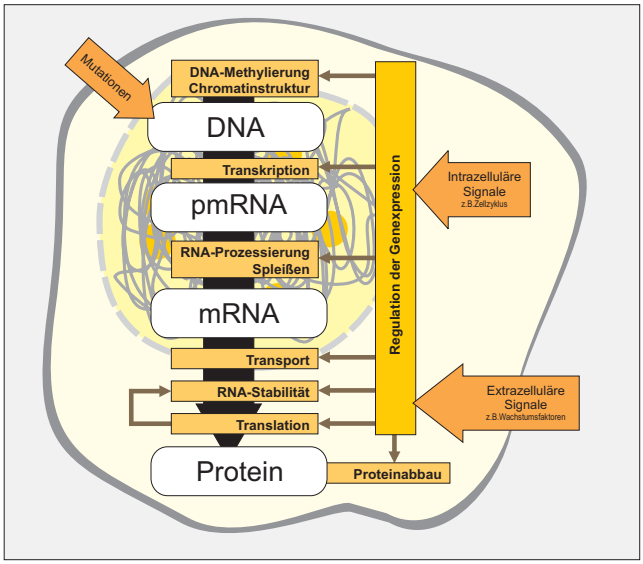


Abbildung 1.3: Regulation der Genexpression bei Eukaryoten

wichtigste Charakteristik ist sicher die Proteinzusammensetzung. Eine Möglichkeit bestünde somit in der quantitativen Bestimmung der Molekülanzahl einer jeden Proteinspezies. Eine weitere Charakteristik ist die Informationsebene, die zwischen Genom und Proteom liegt, nämlich die Gesamtheit aller transkribierten mRNAs (Transkriptom). Obwohl die mRNA-Transkription direkt mit der Proteinproduktion zusammenhängt, ist die Anzahl der vorhandenen mRNA-Moleküle eines Genes nicht proportional zur Molekülanzahl des Genproduktes [Gygi1999]. Diese ist vielmehr eine Näherung für die zeitliche Veränderung desselben. Sie setzt sich zusammen aus dem Normalumsatz ($N_{Aufbau} = N_{Abbau}$) zur Erhaltung des momentanen proteinogenen Zellzustandes und der zeitliche Veränderung desselben ($N_{Aufbau} \neq N_{Abbau}$).

Ein direkter funktioneller Zusammenhang kann deswegen nur zwischen der Anzahl (Zustand) einer RNA-Spezies und der aufbauenden Veränderung der Anzahl des korrespondierenden Genprodukt (Protein) bestehen. Dabei kann, strukturell bedingt oder durch Regelfunktionen der Zelle, die Effizienz der Proteinbildung zwischen einzelnen mRNAs unterschiedlich sein. Die Regelung auf Transkriptionsebene ist wie schon erwähnt die wichtigste (häufigste) Regelung. Einmal fertig prozessierte mRNAs sollten daher vorwiegend durch ihre Stabilität reguliert werden. Die Anzahl der Proteine, die aus einem mRNA-Molekül durchschnittlich entstehen, ist abhängig von dessen Stabilität und von der Affinität zu den Ribosomen. Beide Eigenschaften sind für sequenzidentische mRNAs (gleiche mRNA-Spezies) gleich und ohne weitergehende Regelung zeitlich unverändert. Man kann daher vereinfacht von einer faktoriellen Abhängigkeit der Anzahl gebildeter Proteine von der Anzahl der Moleküle der korrespondierenden mRNA-Spezies ausgehen. (Ein anderer Regulationsmechanismus setzt genau hier an und beeinflußt die Stabilität der einzelnen mRNA-Spezies durch die Bindung von kleinen interferierende RNA-Moleküle - siRNAs)

Eine weitere Form der Differenzierungscharakterisierung existiert auf DNA-Ebene durch die Bestimmung aller aktiven Gene, d.h. alle Gene, die zum Zeitpunkt der Zustandsbestimmung, ohne weitere Veränderung der DNA transkriptierfähig sind. Transkriptierfähig ist ein Gen, wenn sein Mutations-, Methylierungs- und Strukturzustand eine Transkription zu einer translatierbaren RNA-Spezies zu läßt. Das ist zum Beispiel durch Mutations- und Methylierungsanalysen möglich [Adorjan2002].

Wenn man also den Differenzierungszustand einer Zelle möglichst global beschreiben und verstehen will, muß man möglichst alle der Zustandsparameter gleichzeitig bestimmen. Für das Verständnis der

Regulationsmechanismen kommt noch die Erfassung der zeitlichen Veränderung dazu. Momentan kann das weder für eine einzelne Zelle, noch für einen Zellverband vollständig bestimmt werden. Einzelne Proteinformen und RNAs lassen sich durchaus in einzelnen Zellen (*in situ*) qualitativ/ semiquantitativ bestimmen [Lippincott-Schwartz2001] [Perlette2001]. Doch für die Quantifizierung vieler verschiedener Zellmoleküle bedarf es noch immer *in vitro* Methoden, die aber den Aufschluß der Zelle notwendig machen. Man braucht bisher für einen sinnvollen Aufschluß viele Zellen, wodurch man immer einen Durchschnitt über alle vermessenen Zellen erhält. Dafür bekommt man genügend Material um überhaupt bestimmte Moleküle nachweisen zu können.

Auch ist der Expressionszustand bezüglich der Zelluhr zeitlich variabel [Yamaguchi2001], so daß auch Präparationsdauer und -methode einen Einfluß haben. Anschließend zum Aufschluß erfolgt eine Trennung (Chromatographie, Zentrifugation, Filtrierung, Fällung, Magnetbeads ...) und die anschließende Charakterisierung und Quantifizierung. Eine Möglichkeit viele verschiedene Parameter des Zellaufschlusses gleichzeitig zu bestimmen, ist die automatische Parallelisierung der verschiedenen Nachweisreaktionen. Dabei wird entweder der Zellextrakt auf verschiedene Reaktionen (z.B. Mikrotiterplatten) verteilt, oder es werden möglichst viele Reaktionen nebeneinander darin ausgeführt. Um eine möglichst effiziente gleichzeitige Messung zu ermöglichen, bedarf zweiteres einer räumlichen Trennung. Das kann z.B. durch Immobilisierung spezifischer Sondenmoleküle auf einer Trägeroberfläche erfolgen. Ist die Position auf dieser Oberfläche spezifisch für eine Sonde, kann leicht über den Ort, an dem eine Nachweisreaktion stattfindet, die Art derselben zugeordnet werden. Das ist das Prinzip des Sondenarrays. Dieses Prinzip ist generell anwendbar für jeden immobiliserbaren Sondentyp. Zweckmäßig für ein konkretes Array ist die ausschließliche Verwendung von Sonden, die unter den gleichen experimentellen Bedingungen reagieren. Man kann die Arrays daher nach dem Sondentyp (DNA- oder Proteinarray) und dem Sondensubtyp (Antikörper-, Plasmid- oder Oligoarray) unterscheiden. Weitere Unterscheidungen werden nach dem Verwendungszweck (Geneexpressionsarrays, Mutationsarrays, SNP -Arrays), der Arraygröße und Integrationsdichte (Makro- oder Mikroarrays) oder nach dem Trägermaterial (Nylon-, Glas-, Plastikarrays) getroffen. Im folgenden werden nur noch Genexpressionsarrays auf der Basis von DNA-Sonden behandelt, da sie im Moment die am meisten verwendete Form der Bioarrays (auch Biochips) darstellen und die Grundlage für die nachfolgende Arbeit liefern.

1.2 Grundlagen der Arraytechnologie

> „Imagine a 1 cm^2 chessboard. Instead of 64 squares it has thousands, each containing DNA from a specific gene. This is a DNA microarray“ Mark Patterson [Berns2000]

Das grundlegende Prinzip der Expressionsmessung mittels DNA-Arrays ist schon seit längerer Zeit in den Northern Blots bzw. Southern Blots realisiert [Southern1974]. Dabei wird die zu messende Probe (RNA/DNA) nach einer Gelauftrennung auf eine Oberfläche (Nitrozellulose) immobilisiert und mit radioaktiv markierten Sonden hybridisiert. Dabei bindet die Sonde mit bekannter Sequenz spezifisch an die nachzuweisende Probe mit komplementärer Sequenz. An den Stellen des Filters an den die Sonde bindet, kann diese über die radioaktive Markierung (z.B. ^{33}P) nachgewiesen und quantifiziert werden. Die Quantifizierung erfolgt über den Schwärzungsvergleich eines fotografischen Films. Damit lassen sich spezifische Nukleinsäure-Sequenzen zwischen verschiedenen Proben bestimmen und vergleichen.

Die Umkehrung dieses Prinzips wurde dann durch Dotblotting verwirklicht. Hierbei werden spezifische Sonden auf eine Membran (Nitrozellulose/später Nylon) aufgetragen. Die Anordnung dieser Sonden entspricht einem regelmäßigem Gitter (oder Array). Diese Membran wird zusammen mit der radioaktivmarkierten Probe hybridisiert. Nach einem Waschschritt sind die markierten Probenmoleküle nur an ihren korrespondierenden Sonden spezifisch gebunden. Die Auswertung erfolgt wiederum mit Filmen. Der große Vorteil gegenüber den Blot-Methoden ist die Möglichkeit in einem Experiment gleichzeitig mehrere Gene zu erfassen. Zwei Handicaps traten bei dieser Methode auf: Die Verwendung von Radioaktivität und

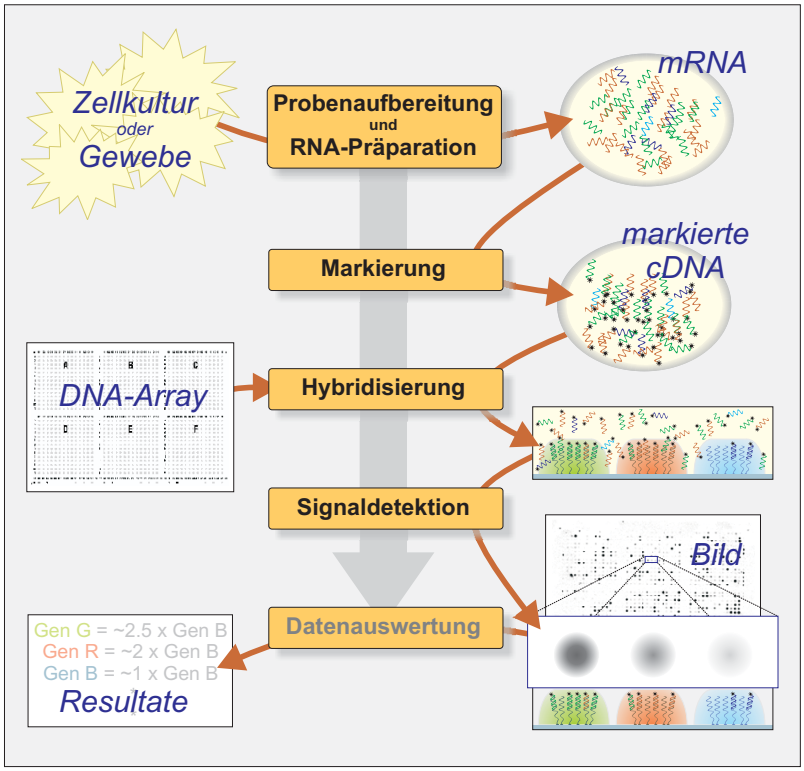


Abbildung 1.4: Schematische Darstellung der Schlüsselschritte bei der Genexpressionsanalyse mit Bioarrays

die riesigen Substanzmengen, die zum Spotten und Hybridisieren erforderlich sind. Um die benötigten Substanzmengen zu reduzieren, projektierte bereits Southern miniaturisierte Formen dieses Verfahrens. Andere Arbeitsgruppen fanden mit der Anwendung chemoluminiszenter Reaktionen einen Ersatz für die radioaktive Markierung [Karger1993].

In der englischsprachigen Literatur herrscht eine gewisse Unklarheit über die Begriffe *probe* und *target* in der Verwendung bei Bioarrays [NatureGenetics1999]. Auch im deutschen Gebrauch sind die entsprechenden Begriffe *Probe* und *Sonde* nicht klar definiert. Für diese Arbeit gilt daher folgende Definition: Eine Probe ist die zu untersuchende, selektierte Form biologischen Materials. Aus ihr leitet sich die Hybridisierungsprobe ab, die auf einen Array zur Untersuchung gegeben wird. Eine Sonde (*probe*) ist eine definierte Form einer Substanz, die geeignet ist eine Zielsubstanz (*target*) in einer Probe selektiv und spezifisch nachzuweisen. Im Falle der Arrays sind es die auf den Array immobilisierten Nachweissubstanzen. Die Bedeutung von *Probe* und *probe* ist damit unterschiedlich.

Einen weiteren Schub in Richtung der Miniaturisierung kam in den neunziger Jahren durch die in situ Synthese von Nukleinsäureoligomeren auf verschieden Oberflächen und mit diversen Methoden (Fotolithografisch [Fodor1991], mittels Siebdruckverfahren [Ermantraut1997]). Andererseits entwickelten sich auch die Auftragungsmethoden (*Spotting/ Printing*) weiter, so daß durch Automatisierung derselben, presynthetisierte Nukleinsäureoligomere, PCR-Produkte oder Plasmide reproduzierbar auf immer kleinere Filter/Chip gespottet werden konnten (Für einen Überblick siehe [Woelfl2000]). Beide Ansätze ermöglichten auch erste kommerzielle Arrayprodukte (z.B. Affymetrix, Clontech), wodurch die Verbreitung der Technik nochmals gesteigert wurde. Für die Detektion der hybridisierten Probe erwies sich

Fluoreszenzmarkierung als geeignet, zumal sie gerade für die Miniaturisierung höhere spatiale Auflösungen ermöglichten, als die damaligen Methoden der Radioaktivitätsbestimmung (Phosphorimagerscreens). Ein Nachteil der Auftragung und Bindung von presynthetisierten Nukleinsäurenoligomeren auf Glasoberflächen war bzw. ist noch immer die hohe Varianz der abgelegten Sondenmenge. Diese führte dazu, daß selbst Ergebnisse von Arrays des gleichen Produktionszeitraumes nur schlecht miteinander vergleichbar sind. Um diesen Nachteil zu umgehen, wurde die Methode der kompetetiven Hybridisierung entwickelt [Schena1996] [DeRisi1997]. Bei dieser hybridisiert man zwei unterschiedlich fluoreszenzmarkierte cDNA-Proben mit dem gleichen Array. Das meßbare Signalverhältnis der beiden unterschiedlich-farbigen Fluoreszenzsignale ist dabei im Idealfall gleich dem molaren Verhältnis der jeweiligen mRNA-Spezies zwischen den beiden Proben (siehe auch Abschnitt 4.1.5). Dieser geniale Schritt erlaubte die relativ billige Herstellung und Verwendung von Glaschips mit maximal etwa 10000 Sonden in vielen wissenschaftlichen Laboren weltweit. Ein großes Problem haben aber alle Auftragungsmethoden. Je größer die Anzahl der zu spottenden Sonden wird, desto schwieriger ist deren Sequenzverifizierung und konsistente Verwaltung, besonders bei klonalen Sonden (Plasmide und die meisten PCR-Produkte). In diesem Sinne haben die auf dem Chip synthetisierten Sonden einen klaren Vorteil [Lipshutz1999], jedoch war ihr Preis bisher relativ hoch.

In den letzten drei Jahren haben dann im Wesentlichen Detailverbesserungen einen merklichen Qualitätssprung der nun vorwiegend kommerziellen DNA-Arrays ergeben. In den Detektionsmethoden erfolgten Verbesserungen in der Stabilität der Fluoreszenzfarbstoffe und im Auflösungsvermögen sowohl der radioaktiven Phosphorimagerplatten und -scanner wie auch der Fluoreszenzscanner. Dazu kommt noch eine densitometrische Methode, bei der die Kinetik der Silberabscheidung an, über Streptavidin/Biotin gebundenen, goldmarkierten Probenmolekülen gemessen wird [Hayat] [Clondiag].

Wie man leicht sieht, ist die Arraytechnik keine einheitliche Methode. Die verschiedenen Entwicklungen koexistieren noch immer. Bisher hat sich noch keine „beste“ Methode herauskristallisiert, da alle ihre Stärken und Schwächen haben. In Tabelle 1.1 sind verschiedene Arraytypen und ihre Eigenschaften zusammengestellt. Ein Quasistandard stellt die Plattform (Oligochips & Meßaparatur) der Firma Affymetrix dar, schon aufgrund ihrer weiten Verbreitung. Allen Methoden gemeinsam sind folgende Schlüsselschritte: Probenaufbereitung $\mapsto$ Markierung $\mapsto$ Hybridisierung $\mapsto$ Signaldetektion $\mapsto$ Auswertung (Schematische Darstellung in Abbildung 1.4).

1.3 Auswertung von Arrayexperimenten

Neben der rein gerätetechnischen Weiterentwicklung haben sich etwas zeitversetzt die Methoden zur Datenbehandlung, -auswertung und -haltung entwickelt. Auch hier entstand eine unüberschaubare Menge an Konzepten, Verfahren und Anpassungen an die jeweilige Meßmethode. Ich werde im nächsten Kapitel auf einige dieser Konzepte eingehen und die einzelnen Schritte der Messung einer genaueren Analyse unterziehen, um ihren Einfluß auf das letztendlich gewonnen Datenmaterial zu bestimmen. Die wesentlichen Schritte der Auswertung sind in Abb. 1.5 dargestellt.

Die Auswertung solcher Daten ist vom Prinzip her gleich der Auswertung von herkömmlichen biologischen Experimenten. Nur die Größenordnung der zu verarbeitenden Daten ist unterschiedlich. Auch dort kann man sie in primäre und sekundäre Auswertung einteilen. Die primäre Auswertung besteht darin, experimentelle Unterschiede in den gewonnen Daten zu erfassen, zu quantifizieren und solche, rein meßtechnischer Art, aus den Daten herauszurechnen (Normalisierung). Dieses Vergleichbarmachen der Daten ist allerdings nur im Rahmen bestimmter statistischer Grenzen möglich.

Daran anschließend erfolgt (wenn möglich) die Bestimmung der statistischen Schwankungsbreite. In diesem primären Block gehen im Idealfall keine biologischen Prämissen ein. Erst danach versucht man in der sekundären Analyse, die gewonnenen Daten durch Hinzunahme biologischer Information zu interpretieren (z.B. Zuordnung von Genexpressionswerten zu metabolischen Karten [Doninger2003]). Wiederum die Menge und Komplexität der Daten macht es erforderlich, daß dazu diverse mathematischer Analy-

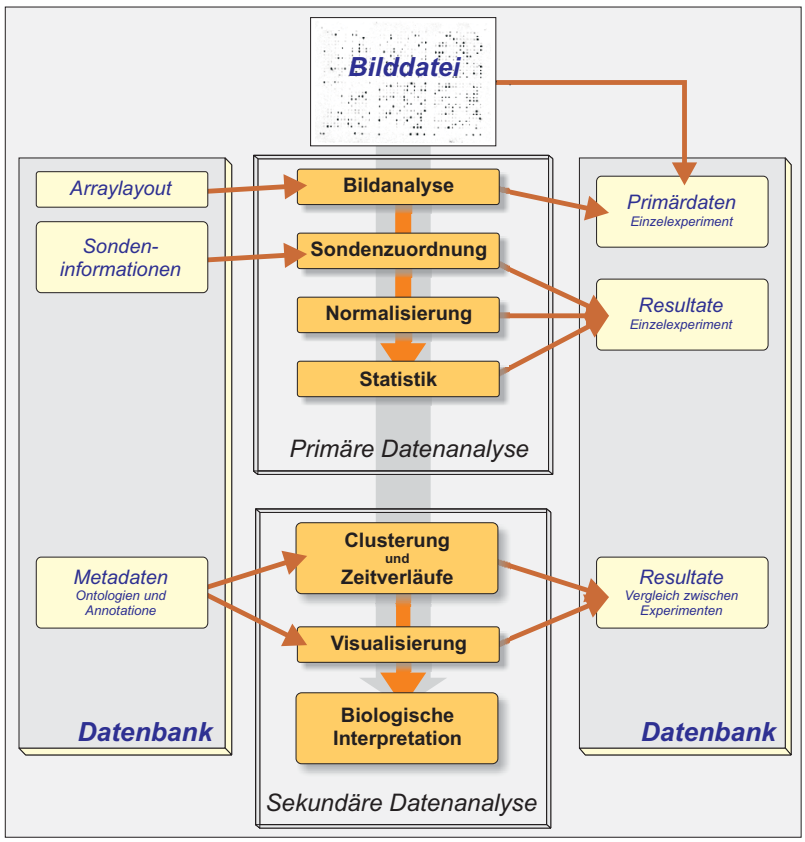


Abbildung 1.5: Schematische Darstellung der Hauptschritte bei der Datenauswertung von Genexpressionsexperimenten

semethoden herangezogen werden müssen (z.B. Bootstrapping Cluster Analysis [Kerr2001A]).

Um die primäre Datenanalyse möglichst frei von biologischen Prämissen zu halten, bedarf es eines durchgehenden Verständnis der Meßmethode. Doch je komplexer die Meßmethode ist, desto mehr (oft nicht quantifizierbare) Einflußfaktoren treten auf. Die arraybasierten Genexpressionsanalyse ist ein solcher Fall. Erschwerend kommt noch die oben beschriebene Vielfalt an Methoden hinzu. Zwei Hauptstrategien zur Normalisierung kommen daher in Betracht: Erstens eine möglichst vollständige Erfassung der Meßfunktion, welche dann nur für eine bestimmte Methode gilt, oder zweitens eine nachträgliche Transformation mittels irgendeiner (vermuteten) inhärenten Eigenschaft der Daten. Das folgenden Kapitel „Systemanalyse“ wird sich näher mit der Problematik der Normalisierung und den Grundlagen der primären Datenanalyse beschäftigen.

1.4 Welche konkreten Fragestellungen lassen sich nun mit diesen Techniken bearbeiten?

Eigentliches Ziel der Genexpressionsmessung ist die Bestimmung der molaren Menge einer jeden mRNA-Spezies in einer bestimmten Zelle zu einem bestimmten Zeitpunkt. Abgesehen von dem immensen informatischen Aufwand die Myriaden dabei entstehender Daten aufzuarbeiten, sind wir noch Jahre von diesem Ziel entfernt. Die Methoden sind bisher noch zu ungenau und zu unvollständig. Doch für die

Typ	**Nylonmakroarrays**		**Nylonmikroarrays**		**Plastik**	**Glasmikroarrays**		**Oligonucleotidchips**	
Subtyp			radioaktiv	colorimetrisch					
Hersteller	allgemein	Clontech	allgemein	allgemein	Clontech	allgemein	Clontech	allgemein	Affymetrix
Array		HAA1.2							H133A
Sonden	klonierte cDNA (Plasmide oder PCR-Produkte)	200- bis 600mere PCR-Produkte	klonierte cDNA (PCR-Produkte)	klonierte cDNA (PCR-Produkte)	synth. 80mere	cDNA Klone &klonierte cDNA (PCR-Produkte)	synth. 80mere	20mere synth. in situ	25mere synth. in situ
Format (Spotanzahl)	$50 - 10000$ auf max. $20 \times 20\ cm^2$	1185 auf $8 \times 12\ cm^2$	200 auf $5 \times 4\ mm^2$	9600 auf $2.7 \times 1.8\ cm^2$	8000 auf $2 \times 5\ cm^2$	$50 \ldots 6400$ auf $1.8 \times 1.8\ cm^2$	3800 auf	64000 auf $1.28 \times 1.28\ cm^2$	500000 (45000 Sondensets)
Trägermaterial	Nylonmembran	Filtermembran	Nylonmembran	Nylonmembran	Plastikoberfläche	Glasträger	Glasoberfläche	diverse	Silizium
Probenmenge	25 μg total RNA	10 μg total RNA oder 2 μg poly A+ RNA	$\approx 0.1\ \mu g$ total RNA	1 μg mRNA	$0.5 \ldots 5\ \mu g$ total RNA oder 1 μg poly-A+RNA	2 μg mRNA	nicht direkt angegeben, OD-Empfehlungen	10 μg mRNA	
Detektionsmethode	radiometrisch $^{33}P/^{32}P$	radiometrisch $^{33}P/^{32}P$	radiometrisch $^{33}P/^{32}P$	colorimetrisch	radiometrisch ^{33}P	fluorometrisch	radio- oder fluorometrisch	fluorometrisch	fluorometrisch
Hybridisierungsvolumen	40 ml	10 ml	100 μl	10 ml	$\approx 15\ ml$	100 μl ... 10 ml	2 ml	200 μl	
Flächendetektor	Phosphorimagerscreen	Phosphorimager oder Autoradiogramm	Phosphorimagerscreen (hochauflösend)	Flachbettscanner	Phosphorimagerscreen	Konfokalscanner	Phosphorimagerscreen oder Floureszenzscanner	Konfokalscanner	Konfokalscanner
Detektionslimit (Anteil mRNA-Moleküle an Gesamtmenge)	1/20000	10-20 Kopien pro Zelle	1/10000	$\approx$ 1/20000	k.A.	1/100000	k.A.	1/300000	1/100000
Minimale Sondenmenge für Detektion	$25 \cdot 10^6$ Moleküle	k.A.	$0.2 \cdot 10^6$ Moleküle	$60 \cdot 10^6$ Moleküle	k.A.	$20 \cdot 10^6$ Moleküle	k.A.	$30 \cdot 10^6$ Moleküle	k.A.
Sequenzüberprüfung	meist teilsequenziert	100% sequenziert	meist teilsequenziert	meist teilsequenziert	100% getestete Oligos	meist teilsequenziert	100% getestete Oligos	systeminherent bestimmt - Abruchsequenzen möglich	systeminherent bestimmt - Abruchsequenzen möglich
Mehrmalige Verwendung	Ja	Ja	Ja	Nicht empfohlen	Ja	Nein	Nein	Nein	Nicht empfohlen
Hybridisierungstemperatur		$68°C$			$60°C$		$50°C$		
Quelle	[Bertucci1999]	[Clontech2000], [Clontech2001]	[Bertucci1999]	[Bertucci1999]	[Clontech2001], [Clontech2002B]	[Bertucci1999]	[Clontech2001], [Clontech2002A]	[Bertucci1999]	[Affymetrix2001]

Tabelle 1.1: Zusammenstellung und Vergleich verschiedener Arrayplattformen

Eierstockkrebs	[Welsh2001]
Ewing's Sarkom	[Welford1998]
Brustkrebs	[Perou1999], [Perou2000], [Hedenfalk2001], [Sorlie2001], [Kroll2002A]
Magenkrebs	[Hippo2002]
Diffuser großer B-Zell Lymphom	[Alizadeh2000]
Lungenkrebs	[Bhattacharjee2001], [Garber2001]
Malignes Melanom der Haut	[Bittner2000]
Schilddrüsenkrebs	[Huang2001]

Tabelle 1.2: Beispiele für krebsassozierte GE-Array-Untersuchungen

meisten Fragestellungen reichen meist Vereinfachungen aus.

So ist das primäre Ziel vieler Experimente weniger die einzelnen Expressionswerte der einzelnen Gene, sondern eher ein Screening nach interessanten Genen. Das sind je nach Fragestellung z.B. Tumormarker, die charakteristisch für einen Tumortyp sind oder sonstige auffällig regulierte Gene, die vielleicht Schlüsselgene in einer bestimmten Krankheitsentwicklung oder allgemein Zellentwicklung sind. Diese könnten als Angriffsziele für Therapien verwendet werden. Für dieses Screening sind einfachere Experimente möglich, bei denen es nur auf die relative Veränderung zwischen den verglichenen Zuständen ankommt. Somit ist es auch mit den billigeren und weniger sensitiven, kompetitiven Arrays möglich dieses Screening durchzuführen. Tatsächlich sind beide Methoden (kompetitive und Einzelhybridisierung) für eine Reihe von Tumoren angewandt worden (Tabelle 1.2). Dieses ist nur ein kleiner Ausschnitt an behandelten Fragestellungen. Eine Literatursuche in der PubMed-Datenbank [PubMed] nach „expression array“ und „cancer / tumour“ ergibt hunderte Treffer. Auch für komplexere Fragestellungen, wie die Untersuchung der zeitlichen Veränderung der Genexpression, lassen sich diese relativen oder semiquantitativen Methoden verwenden. Der letztendliche Erfolg ist umstritten. Je nach Fragestellung reicht die Sensitivität der Methode bei gering exprimierten Genen noch nicht aus, um sichere Aussagen über alle wichtigen Signalproteine treffen zu können, oder das gesuchte biologische Signal ist insgesamt nicht ausreichend, um deutlich aus den immer vorhandenen Variationen der Messung oder der biologischen Proben herauszustechen, so daß kein eindeutiges Schlüsselgen charakterisiert werden kann. Diese Limitierungen werden in Zukunft sicher weiter reduziert, doch ist dazu ein genaues Verständnis aller verwendeten Methoden notwendig.

Kapitel 2

Funktionelle Analyse des Meßsystems und Grundlagen der Normalisierung

> „Wer viel mißt, mißt viel Mist. Was nicht schlimm ist, wenn man weiß, was von dem, was man mißt, Mist ist!“ [1]

Die vielfältigen Variationsmöglichkeiten der Genexpressionsanalyse und ihrer Parameter spiegeln sich natürlich auch in der primären Analyse der gewonnenen Daten wieder. So muß die Analysenmethode die jeweiligen Bedingungen und ihrer möglichen Fehler berücksichtigen. Viele der konkreten Einstellungen an den verwendeten Geräten haben einen Einfluß auf die Daten. Es ist die Aufgabe der Normalisierung diesen Einfluß zu bestimmen und herauszurechnen, damit die Meßunterschiede nicht die biologischen Unterschiede überlagern. Weiterhin muß überprüft werden, ob die Meßunterschiede überhaupt eine Vergleichbarmachung einer Experimentalreihe zu lassen, oder ob extreme Bedingungen zu „irreparablen“ Daten geführt haben. In diesem Sinne hat die primäre Datenauswertung auch immer die Aufgabe einer Qualitätssicherung [Schuchardt2000].

Für die Normalisierung der Genexpressionsdaten sind zwei grundlegende Vorgehensweisen denkbar.

1. ... könnte man die vollständige Signalfunktion bestimmen und so über die Messung der dazu notwendigen Parameter das ursprüngliche Signal (Anzahl einer bestimmten mRNA Spezies) berechnen. $\mapsto$ prospektive Normalisierung

2. ... ließen sich durch Kenntnis unveränderter Charakteristika in den Daten die relativen Meßunterschiede bestimmen und eliminieren. $\mapsto$ retrospektive Normalisierung

Prinzipiell wäre der ersten Methode der Vorzug zu geben, doch ist eben die vollständige Signalfunktion nicht explizit bekannt. Daher müssen verschiedene Annahmen über die einzelnen Schritte der Messung gemacht werden. In den folgenden Abschnitten werden diese kurz funktionell beschrieben und es wird näher auf diese Annahmen eingegangen.

Die zu messende Größe ist die molekulare Menge (Stoffmenge, Molekülanzahl) n_{mRNA} einer bestimmten RNA-Spezies (RNA-Sequenz) in einer Probe. Diese Größe wird durch die Gesamtmessung beeinflußt. Jeder einzelne Schritt kann dabei als Funktion formalisiert werden (Gl.2.1 bis Gl.2.6).

Zusammengesetzt ergibt sich Gleichung 2.8 als allgemeine Signalfunktion. Im Realfall läßt sich meist nur die Meßfunktion M des Flächendetektionssystem vollständig beschreiben. Die anderen Teilfunktionen sind schwerer zugänglich. Die Hybridisierungsfunktion H läßt sich für reale Bioarrays nicht vollständig

[1] Sprichwort

$$\textbf{P}\text{robennahme } (\textit{Sampling}) \Rightarrow \quad n_{sampled} = P(n_{mRNA}) \tag{2.1}$$
$$\textbf{A}\text{ufreinigung} \Rightarrow \quad n_{pur} = A(n_{sampled}) \tag{2.2}$$
$$\text{Markierung } (\textbf{\textit{L}}\textit{abeling}) \Rightarrow \quad \nu_{label} = L(n_{pur}) \tag{2.3}$$
$$\text{Array}\textbf{h}\text{ybridisierung} \Rightarrow \quad \nu_{label,hyb} = H(\nu_{label}) \tag{2.4}$$
$$\text{Markierungs}\textbf{m}\text{essung} \Rightarrow \quad s_{label} = M(\nu_{label,hyb}) \tag{2.5}$$
$$\text{Bild}\textbf{q}\text{uantifizerung} \Rightarrow \quad s_{quant} = Q(s_{label}) \tag{2.6}$$
$$\textbf{N}\text{ormalisierung} \Rightarrow \quad s_n = N(s_{quant}) \tag{2.7}$$

Tabelle 2.1: Allgemeine Übertragungsfunktionen

bestimmen (siehe Abschnitt 2.4). Die Markierungsreaktion läßt sich als sequenzspezifische Amplifizierungsfunktion L auffassen. Ihr Verlauf ist stark von der Methode abhängig und nicht vollständig beschreibbar. Der Einfluß der Bildquantifizierung, beschrieben durch Q, ist einfacher zu beschreiben (siehe quantitative Bildauswertung).

$$s = s_{quant} = S(n) \quad = Q\left(M\left(L\left(H\left(A\left(P\left(n_{mRNA}\right)\right)\right)\right)\right)\right) \tag{2.8}$$
$$n = n_{mRNA} = N(s) \quad = P^{-1}\left(A^{-1}\left(H^{-1}\left(L^{-1}\left(M^{-1}\left(Q^{-1}\left(s_{quant}\right)\right)\right)\right)\right)\right) \tag{2.9}$$

Jeder Analyseschritt kann als eine geschachtelte Unterfunktion formalisiert werden. Im folgenden gehe ich aber nur auf Schritte ein, die eine wesentlichen Einfluß auf den Gesamtprozeß haben. Als wesentlich definiere ich hier Schritte, die einen Einfluß auf die chemische Zusammensetzung der Probe haben, die die Sequenzverteilung der Nukleinsäuren in der Probe oder auf dem Array verändern, oder Schritte, die die Signalverteilung verändern. Hierzu gehören alle Hauptschritte der Genexpressionsanalyse mit DNA-Arrays und das Arraydesign.

2.1 Arraydesign

Ein gutes Arraydesign ist eine notwendige Grundlage für eine erfolgreiche Auswertung der erhobenen Daten. Dabei unterscheidet man physikalisch-chemisches Arraydesign (Spotprozesse, Oberflächenbehandlung etc.) und bioinformatisch-statistisches Design (Sondenauswahl, Mehrfachspots etc.). Es existieren für die technische Gestaltung bereits jetzt viele Kombinationsmöglichkeiten von Auftragungstechniken mit Optimierungen für verschiedene Meßmethoden [Hoheisel1998], [Woelfl2000]. Diese Kombination aus Array und zugehöriger Meßtechnik wird auch Plattform genannt. Das biologisch relevante Design umfaßt die Festlegung der Sondensequenzen für die adressierten Gene und diverser Kontrollen. Hier ist die Kombinations- und Auswahlmöglichkeit noch deutlich größer. Die beiden Hauptkriterien sind: die verfolgte biologische Fragestellung und die verfügbaren finanziellen Mittel. Je qualitativ hochwertiger ein Array ist, desto sicherere Aussagen über die Experimente lassen sich treffen. Mit der Auswahl der Plattform ist dieser Schritt für die meisten Experimentatoren abgeschlossen. Für die Auswertung sind bezüglich des Arraydesigns die Antworten auf folgende Fragen von belang:

- Welche Sequenz hat die abgelegte Sonde?
- Sind mehrere Sequenzen pro Spot vorhanden?
- Wieviel Sondenmaterial ist abgelegt?

- Wieviele Spots gehören zu einer bestimmten Sonde?
- Welche Variationen in der abgelegten Menge sind vorhanden?
- Welche unspezifisches Bindungsverhalten hat das Trägermaterial des Arrays?
- Hat die Immobilisierungsmethode der Sonde (z.B. Linkerlänge) einen Selektionseffekt auf die Probenmoleküle (z.B. Größenausschluß)?
- Hat die Immobilisierungsmethode der Sonde (z.B. UV-Crosslinking) einen Einfluß auf den Zustand der Sonde (z.B. Verhinderung der Rückbindung einer denaturierten Doppelstrangsonde - z.B. gespottete PCR-Produkte)?

Diese Fragen sollten eigentlich alle durch den Arrayhersteller beantwortet werden. Das ist meist nicht der Fall. Die Charakterisierung einer Sonde erfolgt fast immer über eine einzigartigen Identifikationscode (ID) des Herstellers. Zu diesem werden Targetinformationen (Welches Gen soll durch diese Sonde erkannt werden ? - dessen Name(n), Genbank-ID, Kurzname) geliefert. Die dahinter stehende Sequenzinformation der Sonde wird nur selten veröffentlicht und ist höchstens auf Anfrage oder/und gegen Geld erhältlich (Ausnahme Affymetrix). Manchmal ist die Sequenz auch nicht vollständig bekannt und die verwendeten Sonden sind nur ansequenziert. Alle weiteren Informationen zur Qualität und Variabilität der Arrays sind auch auf Anfrage nicht direkt beim Hersteller erhältlich. Nur allgemeine Informationen zum Herstellungsprinzip werden veröffentlicht. Für weit verbreitete Arrays sind Veröffentlichungen von Nutzern vorhanden, die diese Fragen teilweise behandeln [Bertucci1999], [Herwig2001].

Für selbstherzustellende Arrays ist natürlich die Richtung der Fragestellung etwas anders. Welches Design braucht man um später optimale Aussagen treffen zu können [Black2002]? Aber an sich steht jeder der offenen Punkte auch hier für eine zu optimierende Eigenschaft. Doch selbst wenn alle diese Informationen bestimmbar und verfügbar wären, ist es sehr aufwendig diese gesamte Information in der experimentellen Analyse zu berücksichtigen, zumal es für viele der Punkte noch keine Strategie gibt, wie diese in der Analyse zu berücksichtigen sind.

2.2 Probennahme und Aufreinigung

Probennahme und Aufreinigung sind die ersten datenbeeinflussenden Schritte in der Messung. Da ihre Aufgabe genau die Selektion der zumessenden Moleküle aus der Gesamtprobe sind, ist ihr Einfluß sehr groß. Je nach Probentyp (Frischgewebe, eingefrorene Gewebeprobe, Zellkultur) gibt es mehrere Standardprotokolle wie die Proben aufgearbeitet werden müssen. Die hier wichtigste Voraussetzung ist die schnelle Verarbeitung von der Probennahme bis zur cDNA-Präparation/Markierung, da die RNA aufgrund der ubiquitären RNAsen sonst einem raschen Abbau unterliegt. Weiterhin ist eine definierte Behandlung der RNA notwendig, um eben möglichst keine präpararartionsbedingte Unterschiede in die Proben hineinzubringen. Variationen in der Probennahme erzeugen sonst immer Variationen in den resultierenden Datensätzen. Leider läßt sich der Einfluß nicht quantifizieren und oft nicht einmal qualitativ beschreiben. Es ist daher von immenser Bedeutung, daß zu vergleichende Proben mit demselben Protokoll und möglichst vom gleichen Experimentator ausgeführt werden.

Obwohl viele mögliche Fehler benannt werden können, ist eine vollständige Fehlerbetrachtung aus den obigen Gründen nicht möglich. Sich hieraus ergebende Datenunsicherheiten lassen sich nur mit gutem Experimentdesign und ausreichenden Experimentwiederholungen statistisch behandeln [Yang2002B].

2.3 Markierung

Die Markierungsfunktion stellt den Zusammenhang zwischen der Anzahl an signalgebender Markierung (Label) und der Anzahl der jeweiligen Probenmoleküle dar. Dieser Zusammenhang ist entscheidend von

der jeweiligen Markierungsreaktion abhängig. Die verschiedenen Möglichkeiten der Markierungsreaktion sind in Abbildung 2.1 dargestellt.

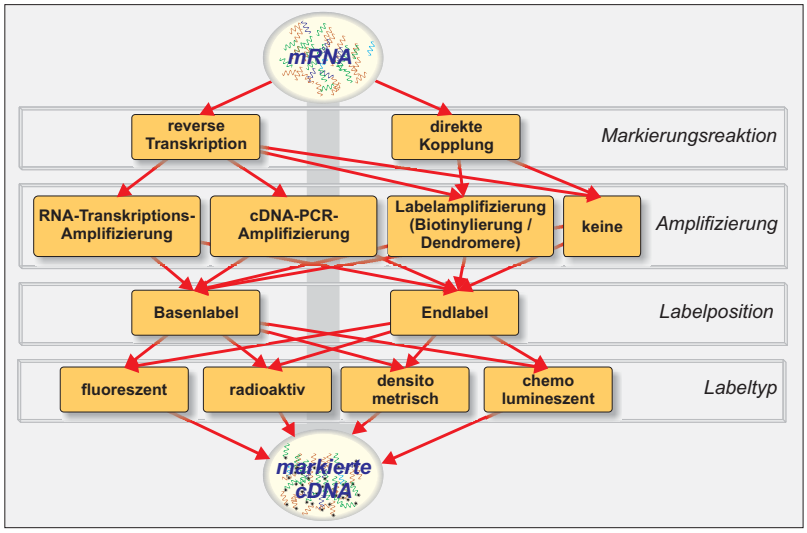


Abbildung 2.1: Kombinationsmöglichkeiten bei der Markierung

Auch wenn im chemischen Sinne das eigentliche Probenmolekül durch den Einbau der Markierungsgruppe (z.B. direkte Endmarkierung mit einem Fluoreszenzfarbstoff) verändert wird oder gar komplett in komplementäre Nukleinsäure umgewandelt wird (z.B. reverse Transkription mit teilweisem Einbau von radioaktiven Nukleotiden), bleibt im Sinne einer molekularen Informationseinheit die Spezifität der Probeninformation erhalten, nur das sie durch den Einbau erst nachweisbar und teilweise auch verstärkt (*amplifiziert*) wird (durch welche Methode auch immer). Bezüglich der Markierungsfunktion wird die Eingangsgröße „Menge an mRNA“ durch die Einbaurate der Markierungsgruppe und die Amplifikationseigenschaften der Reaktion faktoriell in die Ausgangsgröße „Menge an hybridisierbarer Nukleinsäure“ überführt. (Allgemeine Markierungsfunktion Gl. 2.10) Besonders bei intramolekularer Markierung kann die Sequenz der jeweiligen Probespezies einen entscheidenden Einfluß auf die Einbau- bzw. Amplifikationseffizienz haben, so daß es zu Veränderung der Signalverteilung zwischen sequenzunterschiedlichen Probenspezies kommen kann. Bei einer Endmarkierung (Endlabeling) spielt dieses nicht so eine starke Rolle, dafür ist die Amplifikationsmöglichkeiten des Signals (Markierungsrate x Probenanzahl) deutlich geringer.

$$\nu_{label,i} \quad = \quad L(n_{pur,i}, f_{einbau,i}, n_{label}, f_{amp}, V, T) \tag{2.10}$$

Zum Vergleich von Markierungsmethoden für kompetitive Arrays siehe Manduchi et al. [Manduchi2002]. Probleme, die sich durch die Amplifizierung ergeben, werden z.B. von Vernon et al. beschrieben [Vernon2000].

2.4 Hybridisierung

Als Hybridisierung wird im betrachteten Prozeß die Bindungsreaktion zwischen Proben-RNA/DNA und den NA-Sonden bezeichnet. Nach der Markierung werden hiermit die markierten Proben den spezifischen Nachweisplätzen (Immobilisierungsbereiche der jeweiligen Sonden) zugeordnet. Als Signal wird

die letztendliche Anzahl der signalgebenden Moleküle (z.B. radioaktive cDNA) am Nachweisplatz definiert, die nach der Hybridisierung für die Flächendetektion zur Verfügung steht. Diese Reaktion ist von entscheidender Bedeutung für die Spezifität und Selektivität des gesamten Meßprozesses mit NA-Arrays. Grundlage hierfür ist die sequenzspezifische Doppelstrangbildung zwischen Probe und komplementärer Sonde.

$$A + A_c \rightleftharpoons A_c A$$

Die Reaktion ist die Grundlage vieler klassischer Nachweisverfahren wie z.B. PCR und diverse Blotting-Verfahren. Auch dort sind die Einflußparameter der Hybridisierung auf die Daten vielfältig. Wichtige Parameter sind die Zusammensetzung des Reaktionspuffers, der Temperaturverlauf, die Möglichkeit von alternativen Reaktionen und die molaren Verhältnisse aller Reaktanten.

Es haben sich zwei unterschiedliche Varianten der Anwendung entwickelt. Die klassische ist die Einzelprobenhybridisierung, bei der eine einzelne Probe mit einem einzelnen Array hybridisiert wird. Hierbei ist das Hybridisierungssignal die Gesamtmenge an gebundenen Probenmoleküle der jeweiligen genspezifischen Sonde. Die zweite Methode wird kompetitive Hybridisierung genannt. Bei dieser Methode werden zwei unterschiedlich markierte Proben zusammen mit einem einzelnen Array hybridisiert. Die auszuwertende Information ist hier das Mengenverhältnis der cohybridiserten Proben-RNAs an der jeweiligen genspezifischen Sonde (n_{AS}/n_{BS}).

$$A + S \rightleftharpoons AS$$
$$B + S \rightleftharpoons BS$$

Beide Methoden haben ihre Vor- und Nachteile. Es würde den Rahmen dieser erweiterten Einleitung sprengen, hier näher darauf einzugehen, allerdings konnte ich keine Literatur finden, die auf die speziellen Gegebenheiten auf dem Array eingeht, daher habe ich eine ausführlichere theoretische Betrachtung und Simulation dieses Meßschrittes und seiner beiden Varianten vorgenommen (Ergebnisteil S.31).

2.5 Fluoreszenzmessung

Die Meßfunktion des Flächenmeßsystems ist eine gut definierte Funktion, da diese mittels definierter physikalischer Standards empirisch bestimmbar ist. Das Flächenmeßsystem kann je nach Markierungsart verschiedene Detektionsprinzipien verwenden. Hauptsächlich sind das Floureszenzscanner in diversen Ausführungen sowohl für Fluoreszenz- wie auch für radioaktive Markierungen, da die häufig verwendete Radioaktivität meist über Phosphorimager-Platten erfaßt wird und diese wiederum mittels Fluoreszenzmessung ausgelesen werden können. Im Idealfall ist das Meßsignal nur von der Anzahl der auf dem Array vorhandenen Markierungen abhängig und alle anderen Einflußgrößen sind standardisiert und konstant. Im Realfall wird die Messung durch Variationen der Integrationszeit, der Lichtstärke und der Wellenlänge der Anregungslichtquelle (Laser) beeinflußt, sowie durch Schwankungen der optische Eigenschaften der Arrayoberflächen (Sonde, Nichtsonde) und Quantenausbeute durch die Verwendung unterschiedlicher Farbstoffe.

$$s_{label} = M(\nu_{label,hyb}, t, f_q) \tag{2.11}$$

Die Integrationszeit sollte einen linearen Effekt auf die Signalstärke haben. Die Quantenausbeute ist meist für die verwendete Markierung bekannt und kann bei gleicher chemischer Markierungsgruppe als konstant angesehen werden. Der Einfluß der anderen Parameter ist schwerer abzuschätzen. Für die Standardisierung der einzelnen Fluoreszenzmessung und für die Bestimmung der empirischen Meßfunktion lassen sich zum Beispiel Standardarrays mit bekannten Farbstoffmengen verwenden [Kaiser2002] [Clondiag]. (Bei den verwendeten Radioaktivitätsmessungen mittels Phosphorimagerplatten, kam das

interne Kalibrierungssystem der verwendeten Scanner (MD Storm, Fuji FLA5000) zum Einsatz.) Andere Methoden, wie die Silberfärbung, benutzen eine zeitabhängige Meßfunktion. Hierbei wird der Zeitpunkt zum Erreichen einer bestimmten Signalstärke zur Quantifizierung des Meßsignals verwendet. Damit lassen sich Sättigungseffekte umgehen und der dynamische Bereich der quantifizierbaren Werte kann erhöht werden.

Obwohl dieser Schritt sehr gut definiert ist, reicht die pure Kenntnis der Meßfunktion nicht immer aus. Erreicht z.B. die Signalstärke des Fluoreszenzsignals die obere Erfassungsgrenze des Analog-Digital-Wandlers (CCD-chip) kommt es zu einer Signalbeschneidung. Diese läßt keine Rekonstruierung der Werte oberhalb dieser Grenze zu. Diese Signale können nicht normalisiert werden. Aufgrund der Signalverteilung über mehrere Pixel ist dieser Effekt jedoch abgeschwächt. Es gibt keinen harten Schnitt sondern einen sättigungsähnlicher Übergang. Sättigungseffekte können aber auch durch Vorgänge auf dem Array erfolgen (z.B. Hohe Markierungsdichte auf einer Fläche). Ein zusätzlicher nichtlinearer Einfluß ist das Rauschen des Wandlers.

2.6 Bildquantifizierung

Nachdem der Flächendetektor digital ausgelesen wurde, liegt eine Bilddatei zur Auswertung vor. Die Fläche des Detektors ist dabei in Pixel zerlegt. Jeder Pixel enthält dabei 8, 12 oder 16bit-tiefe Signalintensitäten (Graustufen). Da eine Sonde unterschiedliche Flächenausbreitung haben kann, wird eine Hüllkurve definiert (z.B. Kreis), die so groß ist, daß sie jedes reguläre Sondensignal umschließt und somit auf alle Sonden angewandt werden kann. Diese Kurve wird auf die auszumessende Sonde positioniert. Alle Signalwerte der Pixel innerhalb der Hüllkurve (Kreisfläche) werden integriert. Dieses Gesamtsignal enthält noch das Hintergrundsignal des Flächendetektors.

Hintergrundbestimmung Die Hintergrundbestimmung dient zur Messung des Hintergrundsignals des Flächendetektors und des Genarrays. Je nach Meßmethode und je nach Hintergrundqualität können verschiedene Strategien zum Einsatz kommen. Allen Methoden gemein ist die Extrapolation des Hintergrundanteils des Sondensignals aus dem Signal einer bestimmten Referenzfläche. Der Abzug dieses Hintergrundsignals ist nur sinnvoll, wenn sich die Schwankungen im Hintergrund auch physikalisch additiv zum Sondensignal verhalten (siehe Meßrauschen).

1. Globale Referenzmethoden, bei der eine Hintergrund-Referenzfläche ohne Signale für das gesamte Array vermessen wird.
   - Explizit angegebene sondenfreie Spotflächen (global bkg dots -GBD)
   - Globale Hintergrundbestimmung (GIR- „global image region“)
   - Globale Bestimmung der sondenfreien Fläche (MNS - „mode of non spot“)
2. Lokale Methoden, bei denen lokal um jeden einzelnen Sondenbereich ein Hintergrundring oder ein Bereich um ein Subarray vermessen wird, oder sondenfreie Bereiche zwischen den Spots vermessen werden.
   - Lokale-Punkt-Ring-Methode (LDR - „local dot rings“)
   - Lokale-Grid-Ring-Methode (LGR - „local grid rings“)
   - Satelliten-Spot-Methode (SSM - „satellite spots method“)
   - Gewichtete frei definierte sondenfreie Bildflächen (weighted image region -WIR)
   - Gewichtete sondenfreie Spotflächen (weighted bkg dots -WBD)

Die Auswahl der Methode ist immer abhängig von der gegebenen Membran-, Bild- bzw. Experimentqualität [Chen1997]. Dabei sollte möglichst eine Experimentalreihe mit der selben Methode ausgewertet werden. Die globalen Methoden sind dann anwendbar, wenn keine Hintergrundschwankungen sichtbar sind. Das ist selten der Fall, so daß diese kaum verwendet werden. Die lokalen Methoden sind daher die bessere Wahl. Sie haben aber einen entscheidenden Nachteil gegenüber den globalen Methoden. Dadurch, daß bei den lokalen Methoden ein relativ kleine Fläche zur Hintergrundbestimmung verwendet wird, ist der Rauscheinfluß auf den Hintergrundwert viel größer.

2.7 Normalisierung

Die vorangegangenen Abschnitte zeigen, daß die variablen Parameter sehr zahlreich und meist unbestimmt sind und sich somit die explizite Signalfunktion nicht bestimmen läßt. Um dennoch quantitative oder zu mindestens semiquantitative bzw. relative Aussagen treffen zu können, wurden verschiedene Normalisierungsmethoden entwickelt.

2.7.1 Grundannahmen

Alle Normalisierungsmethoden sind abhängig von Ähnlichkeiten zwischen den analysierten Proben und benötigen grundsätzliche Annahmen über das erwartete Verhalten der zu analysierenden Parameter. Die Hauptannahme der Normalisierung ist die funktionelle Abhängigkeit zwischen den biologisch realen Probenunterschieden und den korrespondierenden Werten. Die Eindeutigkeit dieser Funktion ist eine notwendige Voraussetzung für die Normalisierung. Die zweite Annahme ist die Existenz einer unveränderten Eigenschaft innerhalb einer Probe. Diese könnte z.B. die Gesamtmenge an mRNA sein (Globale Methoden), die unveränderte Expression von einem oder einer Gruppe von Haushaltsgenen (Referenzmethode).

2.7.2 Referenzmethoden

Referenzmethoden haben seit jeher eine große Bedeutung für die biologische Quantifizierung. Viele der klassischen Methoden wie Western und Northern Blots nutzen zur Quantifizierung einzelne Referenzgene. Deren Intensitätswerte werden als Skalierungsmaßstab für die eigentlich zu untersuchenden Gene verwendet. Es wird angenommen, daß die **Referenzgene** oder auch Haushaltsgene („housekeeping genes“) unter den jeweiligen experimentellen Bedingungen als gleich exprimiert angesehen werden können, daß heißt, sie werden ständig in der gleichen Menge von den Zellen nachgebildet. Durch die Nutzung von DNA-Arrays und damit möglichen gleichzeitigen Bestimmung verschiedener Haushaltsgene wurden jedoch Zweifel über die Richtigkeit dieser Annahme laut [Adams1993], [Adams1995], [Liew1994], [Spanakis1993].

Eine Verbesserung hinsichtlich des Variationsverhaltens liefert die Verwendung der durchschnittlichen Expression mehrerer Referenzgene [Vandesompele2002].

2.7.3 Lineare Globalisierungsmethoden

Aufgrund der teilweise hohen Varianz der Referenzgene, bezüglich des globalen Verhaltens, wurden Normalisierungsmethoden entwickelt, die auf globalen Kriterien beruhen. Das ist zum einen die Gesamtmenge an gemessenen Signal, welche zur **Mittelwertnormalisierung** führt. Hier liegt die Grundannahmen in:

1. Konstanz der Gesamtmenge an mRNA pro Zelle
2. Proportionalität der hybridisierten mRNA-Menge mit der Gesamtmenge an mRNA
3. Weitgehende Linearität der Signalfunktion

Beide Annahmen müssen durchaus kritisch betrachtet werden. Der erste Punkt kann nur bei biologisch relativ ähnlichen Proben angenommen werden. Der zweite Punkt hat nur Gültigkeit bei nahezu genomischen Arrays und möglicherweise solchen, die eine nicht extreme Genauswahl haben. Arrays mit extremer Genauswahl sind z.B. krankheits- oder funktionsspezifische Biochips, da hier meist nur besonders stark regulierte Gene ausgewählt wurden. Der zweite Kritikpunkt gilt für alle Globalisierungsmethoden (lineare wie nichtlineare).

Die Verwendbarkeit der Methode ist limitiert. Zum einen können bereits einzelne stark regulierte Gene die zweite Grundannahme ungültig machen. Zum anderen wird der Mittelwert durch häufig auftretende Nichtlinearitäten, wie bspw. Sättigungseffekte, beeinflußt.

Um diesem Einfluß zu umgehen, wurden Stutzungsparameter für den Mittelwert eingeführt. Das klassische (symmetrisch) gestutzte Mittel und das in dieser Arbeit beschriebene asymmetrisch gestutzte Mittel [Kroll2002B]. Diese Stutzungen sollen die sättigungsbeeinflußten Bereiche von der Bestimmung des Normalisierungsparameters (Mittelwert) ausschließen.

Der Erfolg der Stutzung ergibt sich aus der prinzipiellen Ähnlichkeit der empirischen Verteilungsfunktion verschiedener Probenhybridisierungen mit einem konkreten Arraytyp. (Eine genauere Erläuterung dieser Ähnlichkeit im Abschnitt 4.2). Eine ähnliche Überlegung führt zur Verwendung des Median (oder anderer Perzentile) als Normalisierungsparameters. (Auch hier erfolgt eine genauere Darstellung im Ergebnisteil 4.4)

Eine andere Herangehensweise ist die Verwendung der linearen Regression zur Bestimmung der Normalisierungsparameters. Dazu wird die Korrelation eines Experimentes zu einem Vergleichsexperiment ausgenutzt. Die Normalisierung mehrerer Experimente ohne ein Referenzexperiment ist durch zyklische Vertauschung der Regressionspaare möglich. Auch hier ist die Normalisierung durch Sättigungseffekte und einzelne extremregulierte Gene beeinflußt. Dieses führte zur Entwicklung der Zentralisierung [Zien2001]. Die Methode benutzt nur einen zentralen Bereich für die Regression. Die Hauptannahme hier ist das Wohlverhalten der Expressionveränderung, d.h. daß die durchschnittliche Veränderung zwischen vielen Proben normalverteilt ist.

2.7.4 Nichtlineare Globalisierungsmethoden

Die linearen Normalisierungsmethoden wurden dahingehend optimiert, daß nichtlineare Effekte einen möglichst geringen Einfluß auf die Normalisierungsparameter haben. Erfolgreich ist die Normalisierung daher auch nur in den Wertebereichen, die selbst nicht durch diese Effekte beeinflußt sind. Die anderen „nichtlinearen“ Bereiche der Werteverteilung werden nur bedingt vergleichbar gemacht.

Betreffen die nichtlinearen Wertebereiche sehr viele Werte oder gibt es gar globale nichtlineare Signalunterschiede, müssen nichtlineare Normalisierungsmethoden angewendet werden. Besonders deutlich wurden solche Effekte bei der Normalisierung von kompetetiven Hybridisierungen. Die unterschiedlichen Fluoreszenzfarbstoffe für die beiden zu vergleichenden Proben führen teilweise zu unterschiedlichen Signalverhalten der beiden Kanäle der Fluoreszenzmessung. Dabei wird meist eine Datentransformation vorgenommen. Das duallogarithmisierte Verhältnis (Logdifferenz) wird gegen das duallogarithmisierte geometrische Mittel (Logsumme) der beiden korrespondierenden Werte aufgetragen [Yang2002A]. Es wird nun angenommen, daß die Veränderungen über den gesamten Wertebereich der Logsumme normalverteilt um die Nichtveränderungsachse liegen sollte (Logdifferenz ist null). Für die Anpassung der Realverteilung an diese Forderung kommt z.B. der laufende Median zum Einsatz. Innerhalb eines laufenden Wertefensters der Logsumme wird der Medianwert der Logänderung auf die Nullachse transformiert.

Ähnlich dazu ist die Methode der lokalen Regression (Loess oder Lowess) [Yang2002A]. Hier wird in einem laufenden Wertefenster eine lineare Regression durchgeführt. Diese Methode kann sowohl an den transformierten Werten ausgeführt werden oder an den Ursprungswerten. Beide Vorgehensweisen führen zu einer Anpassung der Werteverteilung aneinander.

Einen ähnlichen Effekt hat die Angleichung der empirischen Verteilungsfunktion durch Rangwertnormalisierung. Diese Normalisierung wird in Abschnitt 4.4.7 eingeführt und genauer beschrieben. Die

Methode ist ähnlich der Quantilmethode, die Bolstad et. al. [Bolstad2003] beschreiben. Die hier benutzte Methode erweitert die Quantilmethode um eine Fehlerbetrachtung und optimiert die Auswahl des Referenzstandards.

2.7.5 Externe Referenzen und Spiking

Die bisher vorgestellten Normalisierungen sind alle retrospektiv, d.h. sie benutzen vorhandene Daten und versuchen die auftretenden Problem mittels begründeter Annahmen zu behandeln.

Eine andere Strategie wäre, die wichtigsten Schlüsselschritte durch geeignete Kontrollen zu beobachten und somit eine prospektive Normalisierung zu ermöglichen. Die Auswahl geeigneter Standards für jeden Schritt der Genexpressionsanalyse ist dabei sehr aufwendig. Die Kontrollen dürfen nicht mit der Probe wechselwirken und müssen ansonsten die selben Eigenschaften zeigen. Werden die Kontrollen direkt zur Probe dazugemischt werden sie als „Spikes“ bezeichnet. Die mRNA-Spikes bestehen aus verschiedenen Sequenzen, die in unterschiedlicher aber bekannter Menge vorliegen. Nach der Zumischung zur Probe werden die Spike-Sequenzen genauso behandelt wie die der Probe. Der Chip muß nun die entsprechenden Kontrollsonden enthalten. Damit ist es möglich, den gemessenen Signalen die korrespondierenden Stoffmengen der Spikeprobe zuzuordnen. Sind ausreichend viele Spikes unterschiedlicher Menge mitgeführt wurden, kann aus ihren Signalen die Signalfunktion bestimmt werden. Dadurch werden viele der zuvor notwendigen Normalisierungsannahmen überflüssig.

Allerdings ist die konsistente Verwendung von Spikes sehr arbeitsaufwendig und es löst das Normalisierungsproblem nicht vollständig. Das Verhältnis zwischen Probe und Spike kann durchaus varieren und muß retrospektiv korrigiert werden. Eine der wenigen Veröffentlichungen, die diesen Methode verwendet haben, stammt von Van der Peppel et al.[Peppel2003]. Es konnte gezeigt werden, daß die Methode im Detail andere Ergebnisse liefert als retrospektive Normalisierungsmethoden.

Die meisten der bisher generierten Daten wurden jedoch ohne diese Art der Kontrolle generiert und aus Kostengründen wird es auch noch einige Zeit dauern bis sich die durchgängige Verwendung ausreichender externer Referenzen durchsetzt. Die in dieser Arbeit entwickelten Methoden werden daher auch noch einige Zeit ihre Berechtigung haben.

2.8 Meßfehler

2.8.1 Statistische Fehler

Jede Messung zeigt eine gewisse Variabilität bei der Bestimmung eines Signals. Diese Variabilität kann mehrere Ursachen haben. Zum einen ist bei biologischen Proben eine exakt gleiche Probennahme unmöglich, zum anderen zeigt das eigentliche Meßsystem durch variierende Bedingungen selbst ein gewisses Rauschen. Letzteres wird besonders deutlich durch das Auftauchen von negativen Werten bei den hintergrundkorrigierten Signalen. Das macht biologisch, wie chemisch (Hybridisierung) keinen Sinn, und kann nur durch Variationen in der Signalbestimmung verursacht sein. Diese werden zum einen durch echtes Rauschen des Detektionssystem (z.B. Pixelrauschen) hervorgerufen. Weiterhin kommen hierin Fehler in der Hintergrundbestimmung zum Ausdruck. Ursachen sind dafür Fehlsignale innerhalb der Bestimmungsflächen für das Hintergrundsignal durch unspezifische Bindung signalgebender Moleküle in der Probenlösung und durch breite Sockelintensitäten starker Sondensignale (Überstrahlung und Restsondenmaterial in den Außenbereichen der eigentlichen Probe), sowie nicht optimale manuelle/automatische Positionierung. (Zum Problem der Überstrahlung/Sockelintensitäten siehe Machl et. al. [Machl2002]) Die Variation durch Fehlsignale letzterer Art ist schwer zu bestimmen, eine grobe Abschätzung ist aber möglich. Sie hat nur Sinn, wenn lokale Methoden für die Hintergrundbestimmung verwendet wurden. Auch hier nehmen wir als Arbeitshypothse eine Normalverteilung dieses Fehlers an, bzw. daß die Wahrscheinlichkeit, daß ein unspezifisches Signal auftritt, unabhängig von der Position auf der Membran ist. Die

Auswirkungen dieser Variation ist, vorwiegend bei Sonden für nicht- und niedrigexprimierte Gene, durch das Auftreten von bereits erwähnten Negativwerten zu beobachten.

Zu diesem additiven absoluten Rauschen des Gesamtmeßsystems kann noch ein relativer Fehler durch Variationen in der Probennahme, der Sondenmenge auf dem Array, der Hybridisierungszeit usw. hinzukommen. Von Rocke und Durbin wurde dazu ein allgemeines Fehlermodell in die GEA-Analyse eingeführt. Dieses geht von den oben genannten Fehlertypen aus und es wird postuliert, daß über einer „großen“ Anzahl von Experimenten, sich diese Fehler auch normalverteilt verhalten [Rocke2001]. Für die Bewertung von Ergebnissen der GEDA ist die Abschätzung dieser Fehler von entscheidender Bedeutung. Diese Arbeit stellt eine Methode vor, wie der Absolutanteil des Rauschfehlers für Daten von gespotteten Arrays (HAA1.2) abgeschätzt werden kann (Abschnitt 4.3).

2.8.2 Andere Fehlereinflüsse

Negative Werte können auch noch weitere Ursachen haben. Zum Beispiel erscheinen bei manchen glasbasierten Arrays Bereiche mit Hintergrundsignalen, die systematisch höher sind als die in ihnen liegenden Sondenbereiche. Ursache ist hier wahrscheinlich eine unspezifische Bindung signalaktiver (markierter) Moleküle an nicht deaktivierte reaktive Bereiche der Glasoberfläche. Nur die durch die Bindung der Sonde abgesättigten Bereiche sind deaktiviert und zeigen kaum unspezifische Bindung. In diesem Fall ist das Hintergrundsignal nicht additiv zum Sondensignal, die oben aufgeführten Methoden würden versagen.

Weitere Einflüsse sind Bildartefakte, wie Kratzer, Flecken oder Fehlspots, die zwar zufällig auftreten können, aber bezüglich eines Einzelexperiments nicht normalverteilt und somit nicht mit obiger Methode zu erfassen sind. Diese Fehler führen oft zu extremen Abweichungen vom (vermuteten) Realwert. Sie sind bei manueller Bildbearbeitung oft auffällig, können aber bei sehr großen Arrays (>2000 Spots) auch leicht übersehen werden. Da hier der Realwert nicht rekonstruiert werden kann müssen diese Bereiche von der weiteren Analyse ausgeschlossen werden und als Fehlwerte markiert werden. Nur Mehrfachauftragungen einer Sonde auf verschiedene Bereiche des Arrays oder Experimentwiederholungen können hier einem Datenverlust (missing value) für einzelne Sonden („Gene“) vorbeugen.

Außer den rein experimentellen Fehlern hat auch die Normalisierung einen Effekt auf den Fehler der Auswertung. Abschnitt 4.4 zeigt, in wie weit die experimentellen Fehler durch verschiedene Normalisierungsmethoden verstärkt werden können.

Kapitel 3

Material und Methoden

3.1 Materialien

3.1.1 Verwendete Geräte

Auslesegeräte Für die Flächenausmessung der radioaktiven Membranen kamen folgende Geräte zum Einsatz:

- Molecular Dynamics Storm
- Fuji FLA 5000

Computer Für die Datenauswertung wurden normale PCs mit Athlonprozessoren >500MHz und >256MByte Speicher eingesetzt.

3.1.2 Software

Bildauswertung und Spotquantifizierung

- Raytest AIDA array analyzer 3.10 /3.20
- Affymetrix Microarray Suite MAS5.0

Normalisierung, Hintergrundkorrektur, Visualisierung

- Microsoft Excel 97
- Mathworks MatLab 6.0 mit Statistik-Toolbox

3.1.3 Genexpressionsarrays

- *Clontech Human Atlas Array 1.2 - HAA1.2* Die Nylonmembranen der Firma Clontech (jetzt BD Bioscience) enthält 1185 cDNA-Sonden aus einer relativ allgemeinen Genauswahl.
- *Affymetrixarray Hu133A Oligoarray* Der Oligoarray Hu133A der Firma Affymetrix ist ein genomisches Array mit Sondensets aus ≈ 30 20mer DNA-Oligomeren für 22283 Zielgene.

3.1.4 Verwendete Daten

Die in der Arbeit verwendeten Beispieldaten wurden, wenn nicht anders angegeben, in der Arbeitsgruppe von PD Dr.Stefan Wölfl generiert. Es handelt sich um Experimente für diverse biologische Fragestellungen.

Lungendaten Die Daten der Lungenproben stammen von einer Untersuchungsreihe zu Bronchialkarzinom. Bei ursprünglichen Proben handelt es sich um Tumorgewebe und tumorfreie Resektionsränder von operativen Tumorentfernungen. Die Proben wurden von Dr. Jörg Sänger zur Verfügung gestellt (Institut für Pathologie, Bad Berka). Die Probenaufbereitung und Genexpressionsanalyse mit dem Membranarray HAA1.2 wurde von Dr. Larissa Pusch (AG Wölfl, Klinik für Innere Medizin, Universität Jena) durchgeführt. Ich erhielt die generierten Bilddateien. Diese wurden von mir mit dem Raytest-Programm ausgewertet. Bei den bearbeiteten Tumorentitäten handelt es sich um mutmaßliche nicht kleinzellige primäre Bronchialkarzinome. Ein Teil der Bilder wurde noch mit einer älteren Softwareversion von Molecular Dynamics „ArrayVision“ verarbeitet, diese wurden wenn möglich neuverarbeitet.

Nierendaten Die Daten der Nierenproben stammen von einer Untersuchungsreihe zu Nierentumore. Bei ursprünglichen Proben handelt es sich um Tumorgewebe und tumorfreie Resektionsränder von operativen Tumorentfernungen. Die Proben wurden von Dr. Kerstin Junker zur Verfügung gestellt (Institut für Urologie, Bad Berka). Die Probenaufbereitung und Genexpressionsanalyse mit dem Membranarray HAA1.2 wurde von Larissa Pusch durchgeführt. Ich erhielt die generierten Bilddateien. Diese wurden von mir mit dem Raytest-Programm ausgewertet.

Leukämiedaten Die Daten der Leukämieproben stammen von zwei unterschiedlichen Untersuchungsreihen zur Wirkung zweier proteinischer Wirkstoffe auf leukämische Patienten. Die Proben wurden von Dr. Markus Ritter, Dr. Andreas Borchardt und Prof.Dr. Andreas Neubauer (Klinik für Hämatologie und Onkologie, Universität Marburg) zu Verfügung gestellt. Die Probenaufbereitung und Genexpressionsanalyse mit dem Membranarray HAA1.2 wurde von Larissa Pusch (AG Wölfl, KIM 1, Jena) durchgeführt. Ich erhielt die generierten Bilddateien. Diese wurden von mir und von Frau Pusch mit dem Raytest-Programm ausgewertet.

Hefedaten Die in der Diskussion verwendeten Bilder von Filterhybridisierungen stammen von Experimenten zur Genotoxizität von verschiedenen Substanzen auf Hefe. Die Experimente wurden von Frau Ana Kitanovic (AG Wölfl, KIM 1, Jena) durchgeführt. Der Hefearray stammt aus dem Labor von Herrn Jörg Hoheisel und Frau Nicole Hauser (DKFZ Heidelberg)

K562-Daten Die in der Diskussion verwendeten Bilder von Hybridisierungen von einem Affymetrixarray Hu133A stammen von Experimenten einer Zellkulturreihe von K562-Zellen. Die Experimente wurden von Herrn Stefan Knoth (AG Wölfl, KIM 1, Jena) durchgeführt. Die primäre Datenanalyse wurde von mir mit dem Affymetrix MAS5.0 gemacht.

3.2 Hybridisierungsmodelle

3.2.1 Berechnung der Bindungskonstanten

Zur Berechnung der Bindungskonstanten der verschiedenen einzelsträngigen Proben an die komplementären DNA-Sonden wurde der Algorithmus von Breslauer zu Grunde gelegt [Breslauer1986]. Dieser basiert auf dem Nearest-NeighborModell, das davon ausgeht, daß die Bindungsenergien der gebildeten Doppelstränge im wesentlichen von der Stapelenergien (stacking energies) benachbarter Basenpaare

abhängig ist. Die wesentliche Gleichung 4.4 finden sich in Abschnitt 4.1.2. Die notwendigen Parameter wurden aus verschieden neueren Veröffentlichungen entnommen, die verbesserte Parameterwerte enthalten [Sugimoto1996] [SantaLucia1996] [SantaLucia1998] [Allawi1997] [Allawi1998A] [Allawi1998B] [Allawi1998C] [Peyret1999]. Der verwendete Algorithmus zählt die Anzahl der verschiedenen 16 möglichen benachbarten Basenkombination (N_{AA},N_{AC},...,N_{TT}) in einer vorgegebenen Sequenz. Diese Anzahl wird mit dem jeweiligen Entropie, Enthalpie oder freie Energie- Parameter (H_{NN},S_{NN},G_{NN}) der Basenkombination multipliziert. Nun wird jeweils die Summe der jeweiligen Produkte gebildet. Für die Enthalpie kommt zusätzlich noch ein Initiationsterm dazu (H_{init}). Für die Entropie und Energie kommt zusätzlich zu einen Initiationsterm (S_{init} bzw.G_{init}) noch ein Symmetrieterm (S_{sym} bzw.G_{sym}) dazu, falls die zu berechnende Sequenz selbstkomplementär ist. Die Grundgleichungen können auch noch für die Abhängigkeit der Reaktion von der Natriumionenkonzentration erweitert werden [SantaLucia1998].

3.3 Normalisierung und Fehlerbehandlung

3.3.1 Generierung der Testdaten

Virtuelle Primärdaten

Lineare Verteilung:

$$\begin{aligned} r &= 1\ bis\ 1000 \\ I_{S_r} &= 1000 \cdot (1 - \frac{r}{N}) \end{aligned} \tag{3.1}$$

Exponentielle Verteilung:

$$I_{S_r} = 1000 \cdot e^{-10\frac{r}{N}} \tag{3.2}$$

Exponentielle Verteilung modifiziert für Vergleichsarray: Es werden zwei Arrays generiert. 75% der Werte sollen auf beiden Arrays zufällig auf die gleichen Spotpositionen verteilt werden. Diese Werte haben eine logarithmische Werteverteilung. 25% (15×15) der Spots sollen Veränderungen repräsentieren. Dazu werden 15 Werte der gleichen Werteverteilung mit einander kombiniert, so daß 225 Datenpaare entstehen. Diese 225 Datenpaare werden wieder zufällig den restlichen Spotpositionen zugeordnet. Dabei wird dem Array 1 der Wert des erste Element des jeweiligen Datenpaares zugeordnet und dem Array 2 der des zweiten Elements. Aufgrund des Vorkommens von 15 gleichen Datenpaaren sind letztendlich bei etwa 80% der Spotpositionen die Werte zwischen beiden Arrays gleich (690 Spots). Die Rang-Intensitäts-Kurven sind durch dieses Verfahren bei beiden Arrays gleich.

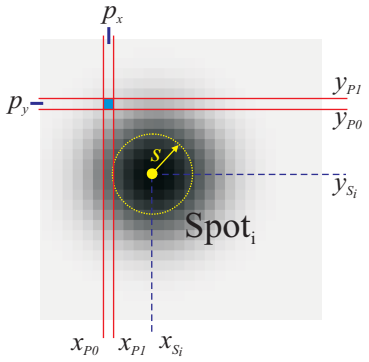
$$75\%\ \text{der Daten(kanalunabhaengig}\ N = 675\ NSpots): \tag{3.3}$$
$$I_{S_r} = 1000 \cdot 2^{-10\frac{r}{N}}$$
$$25\%\ \text{der Daten(kanalabhaengig}\ N = 15\ N^2Spots): \tag{3.4}$$
$$I_{S_r} = 1000 \cdot 2^{-10\frac{r}{N}}$$

I_{s_r}	Gesamtintensität des Spots mit dem Rang r
r	Rang r

Virtuelle Arrays

Die zum Testen der Bildauswerteparameter und Rauschbestimmung benutzten Arraybilder wurden durch eigene MatLab-Funktionen erzeugt. Auf die 1000x1000 Pixel großen Bilder wurde eine virtuelles Grid gelegt. An jeder Gridposition befindet sich ein Spot mit der Gesamtintensität I_s. Diese Intensität ist als

Näherung realer Spots durch eine 2D-Gaußfunktion über das Bild verteilt. Jeder Pixel ist eine Fläche von 1x1 Längeneinheiten. Die Intensität eines Pixel I_p ist das Flächenintegral der Gaußfunktionen aller Spots für die Fläche, die dieser Pixel einnimmt. Für eine Intensität an einer bestimmten Pixelposition gilt Formel 3.5.



$$I_P(x_P;y_P) = \sum_{i=1}^{N} \left(\frac{1}{4} I_{S_i} \left(\begin{array}{c} erf\left(\frac{\sqrt{2}}{2s}\left(x_{P0}-x_{s_i}\right)\right)\cdot erf\left(\frac{\sqrt{2}}{2s}\left(y_{P0}-y_{s_i}\right)\right)+\ldots \\ erf\left(\frac{\sqrt{2}}{2s}\left(x_{P1}-x_{s_i}\right)\right)\cdot erf\left(\frac{\sqrt{2}}{2s}\left(y_{P1}-y_{s_i}\right)\right)-\ldots \\ erf\left(\frac{\sqrt{2}}{2s}\left(x_{P0}-x_{s_i}\right)\right)\cdot erf\left(\frac{\sqrt{2}}{2s}\left(y_{P1}-y_{s_i}\right)\right)-\ldots \\ erf\left(\frac{\sqrt{2}}{2s}\left(x_{P1}-x_{s_i}\right)\right)\cdot erf\left(\frac{\sqrt{2}}{2s}\left(y_{P0}-y_{s_i}\right)\right) \end{array} \right) \right) \tag{3.5}$$

N	Anzahl der Spots
I_{s_i}	Gesamtintensität des Spots i
erf	Errorfunktion
x_i, y_i	Position des Spotmaximums (Spotmitte)
p_x, p_y	Pixelkoordinaten
x_{p0}, y_{p0}	Anfangsposition des Pixels
x_{p1}, y_{p1}	Endposition des Pixels
s	Halbwertsradius der Gaußfunktion

Für die Intensitätsverteilung der Gesamtintensität der 900 Spots wurden drei Verteilungsfunktionen angewandt. Diese Gesamtintensitäten wurden mittels Permutation zufällig über das Positionierungsgitter verteilt. Weitere Modifikationen sind die Addition von Pixelrauschen (normalverteilte Zufallsfunktion) ,die Addition eines Hintergrundgradienten sowie die Simulation der Sättigung mittels einer Wurzelfunktion.

3.3.2 Bestimmung des Sondensignals

Nachdem der Flächendetektor digital ausgelesen wurde, liegt eine Bilddatei zur Auswertung vor. Die Fläche des Detektors ist dabei in Pixel zerlegt. Jeder Pixel enthält dabei 8, 12 oder 16bit-tiefe Signalintensitäten (Graustufen). Da eine Sonde unterschiedliche Flächenausbreitung haben kann, wird eine Hüllkurve definiert (z.B. Kreis), die so groß ist, daß sie jedes reguläre Sondensignal umschließt und somit auf alle Sonden angewandt werden kann. Diese Kurve wird auf die auszumessende Sonde positioniert. Alle Signalwerte der Pixel innerhalb der Hüllkurve (Kreisfläche) werden integriert. Dieses Gesamtsignal enthält noch das Hintergrundsignal des Flächendetektors.

$$I_{Total,Spot} = \sum_{i=1}^{N_{Pixel,Spot}} I_{Pixel,i} \tag{3.6}$$

Verschiedene Methoden sind im „AIDA-array analyzer“ verfügbar, um die Qualität der Gesamtsignals einer Sonde zu bestimmen. Eine wichtige ist die angenommene radiale Gleichverteilung. Dazu wird die jeweilige Intergrationsfläche in vier Kreissektoren unterteilt. Varieren die Signale der Sektoren zu stark voneinander, ist ein Fehler auf dem Filter oder eine falsche Positionierung der Quantifizierungsflächen auf dem Detektorbild anzunehmen. Kann eine Falschpositionierung ausgeschlossen werden muß der Spot als Fehlspot markiert werden. Seine Quantifizierung würde kein sinnvolles Ergebnis liefern [YangMC2001], [Raytest2002AAM].

3.3.3 Hintergrundbestimmung

Die Hintergrundbestimmung erfolgt im allgemeinen mittels einer Referenzfläche. Von dieser wird die Gesamtintensität bestimmt. Über die Anzahl der Pixel der Referenzfläche und der Spotbestimmungsfläche wird mit Gl. 3.7 der Hintergrund der Spotbestimmungsfläche geschätzt.

$$I_{Hintergrund,Spot} = \frac{N_{Pixel,Spot}}{N_{Pixel,Hintergrundflche}} \cdot \sum_{i=1}^{N_{Pixel,Hintergrundflche}} I_{Pixel,i} \tag{3.7}$$

Die verschiedenen nachfolgenden Hintergrundmethoden unterscheiden sich im wesentlichen durch die Definition der Referenzfläche.

Globale Hintergrundbestimmung (GIR- „global image region“)

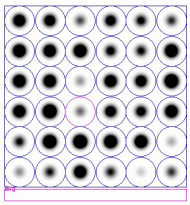


Abbildung 3.1: Globale Hintergrundbestimmung - Integration des Hintergrundes innerhalb des lilanen Rechtecks

Es wird eine (oder mehrere) repräsentative sondenfreie Fläche des Meßbildes vermessen. Der durchschnittliche Hintergrundwert dieser Fläche wird als Hintergrundsignal für alle vermessenen Sondenspots verwendet (Abb.3.1).

Globale Bestimmung der Sondenfreien Fläche (MNS - „mode of non spot“)

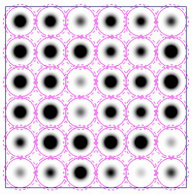


Abbildung 3.2: Globale Bestimmung der Sondenfreien Fläche - Integration des Gesamtbereiches außerhalb der gestrichelten lilanen Kreise

Es wird die gesamte Fläche integriert, die nicht zur Spotmessung verwendet wird, minus einen „Sicherheitsbereich“ der zusätzlich um die jeweiligen Spots gelegt wird (3.2). Der mittlere Pixelwert dieser Fläche wird für alle vermessenen Sondenspots als Hintergrundwert verwendet.

Hintergrundspots

Es werden eine Reihe repräsentativer sondenfreier Referenzspots auf dem Array vermessen. Der Mittelwert der Signalwerte dieser Referenzspots wird als Hintergrundsignal für alle vermessenen Sondenspots verwendet. Diese Hintergrundreferenzspots müssen bereits im Arraydesign berücksichtigt worden sein.

Lokale-Punkt-Ring-Methode (LDR - „local dot rings“)

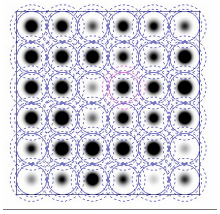


Abbildung 3.3: Lokale-Punkt-Ring-Methode - Integration des Hintergrundes innerhalb der beiden gestrichelten Ringe um einen Spot

Eine von Raytest Aida Image Analyzer angebotene Methode zur Bestimmung eines lokalen Hintergrundes. Dazu wird um die Integrationsfläche des Sondensignals zwei Ringe gelegt. Die Fläche des inneren Ringes wird nicht integriert. Sie dient als Pufferzone gegen Sockelintensitäten des Sondensignals. Das Signal Innerhalb des äußeren Ringes wird integriert und durch eingeschlossene Fläche (Pixelanzahl) dividiert. Dieses Hintergrundsignal gilt nur lokal für den umringten Sondenspot.(Abbildung 3.3)

Lokale-Grid-Ring-Methode (LGR - „local grid rings“)

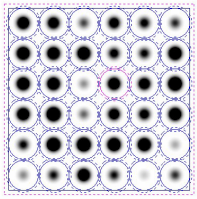


Abbildung 3.4: Lokale-Grid-Ring-Methode - Integration des Hintergrundes innerhalb des gestrichelten Rechtecks und außerhalb der gestrichelten Kreise

Eine von Raytest Aida Image Analyzer angebotene Methode zur Bestimmung eines lokalen Hintergrundes. Dazu wird um eine Untereinheit (Subgrid) des Auswertemusters (Array) zwei „Ringe“ gelegt. Die Fläche des inneren Ringes ist eine Erweiterung der Fläche dieser Untereinheit und wird nicht integriert. Sie dient als Pufferzone gegen Sockelintensitäten der Sondensignals. Das Signal innerhalb der zweiten Erweiterung (äußeren Ringes) wird integriert und durch eingeschlossene Fläche (Pixelanzahl) dividiert. Dieses Hintergrundsignal gilt nur lokal für den umringten Subgrid. (Abbildung 3.4)

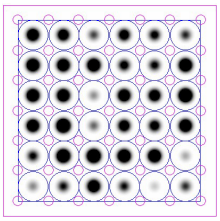


Abbildung 3.5: Satelliten-Spot-Methode - Integration des Hintergrundes innerhalb der magenten Kreise (Nicht direkt in AIDA enthalten) siehe Text.

Satelliten-Spot-Methode (SSM - „satellite spots method“)

Eine weitere Möglichkeit der Bestimmung des lokalen Hintergrundes sind „Satelliten Spots“. Hierzu werden vier kleine Kreisbereiche so um jeden Meßspot gesetzt, daß diese möglichst nahe am Meßspot liegen, aber außerhalb dessen Sockelbereiches und außerhalb der Signale der umliegenden Sonden. Dieses Verfahren ist relativ aufwendig. Es läßt sich aber automatisieren, wenn für jeden konstant distanten $r \times c$ Subarray ein $(r+1) \times (c+1)$ großer Array für die Satellitenspots definiert wird. Dieser wird so positioniert, daß jeder Meßspot von vier Satellitenspots umringt ist. Jeder der inneren Satellitenspots ist dann zu jeweils 4 Meßspots gehörig. Nach der automatische Positionierung kann noch eine Überprüfung der einzelnen Spots vorgenommen werden. Dieses Hintergrundsignal gilt nur lokal für den umringten Sondenspot. (Abbildung 3.5)

3.3.4 Rangbestimmung

Relativ unabhängig von unterschiedlichen Meßfunktionen sind Veränderungen in der Rangfolge der Signale verschiedener Sonden zwischen Experimenten. Dazu wird jedem Element der Meßwertmenge eines Experimentes eine Rangzahl zugeordnet. Man unterscheidet absteigende und aufsteigende Ränge. In dieser Arbeit werden nur absteigende Ränge verwendet, d.h. alle Elemente (Sondensignale) eines Experimentes werden ihrer Größe nach absteigend geordnet. Es gibt genau N Ränge, wobei N die Gesamtanzahl der Elemente (Sonden) ist. Der Rangzahl $r = 1$ wird das größte Element zugeordnet, der Rangzahl $r = 2$ das zweitgrößte , usw., und der Rangzahl $r = N$ das kleinste Element. Falls es mehrere gleichwertige Elemente wird ihr Wert der gleichen Anzahl an Rängen zugeordnet. Die Zuordnung der gleichwertigen Elemente zu den einzelnen Rängen im zugehörigen Rangbereich ist dabei zufällig. Jedem Element wird also genau ein Rang zugewiesen. Dadurch kann jedem Rang ein Wert zu geordnet werden, aber nicht jedem Wert ein Rang, sondern ein Rangbereich, falls mehrere gleichwertige Elemente existieren. Diese eindeutige Zuordnung eines Wertes zu einem gegebenen Rang nenne ich Rangwertfunktion (rgw) mit der folgenden Definition:

$$\mathrm{rgw}(r) \stackrel{def.}{=} \overline{\{x_r | (\mathrm{card}\{x | x > x_r\} < r) \wedge (\mathrm{card}\{x | x < x_r\} < (N - r))\}} \tag{3.8}$$

Der Rangwert des Ranges r ist also der Wert eines Elementes für den gilt, es gibt höchstens r-1 Elemente x die größer sind als dieser Wert und höchstens (N-r-1) Elemente die kleiner sind. Falls mehrere gleichwertige Elemente existieren ergibt die Mengenfunktion die Menge dieser Elemente. Damit nur ein Wert übergeben wird ist der Rangwert der Mittelwert dieser Menge und damit der Einzelwert.

Die Rangfunktion ist liefert für jeden Wert eines Elementes den zugehörigen Rangbereich. Für jeden einzigartigen Wert liefert sie eine Rangzahl. Für n Elemente gleichen Wertes liefert sie n Rangzahlen. Ich definiere die Rangfunktion als:

$$\mathrm{rfun}(x_r) \overset{def.}{=} \begin{cases} \mathrm{card}\{x|x = x_r\} = 1 \rightarrow & \{(\mathrm{card}\{x|x > x_r\} + 1)\} \\ \mathrm{card}\{x|x = x_r\} > 1 \rightarrow & \{(\mathrm{card}\{x|x > x_r\} + 1), \ldots, (N - card\{x|x < x_r\} - 1)\} \end{cases} \tag{3.9}$$

Um eine eindeutige Rückzuordnung zu definieren, werden Rangbindungen eingeführt. Diese Rangbindungsfunktion (rgb) definiere ich als:

$$\mathrm{rgb}(x_i) \overset{def.}{=} \frac{2 \cdot \mathrm{card}\{x|x \in x_{1\ldots N} \wedge x > x_i\} + \mathrm{card}\{x|x \in x_{1\ldots N} \wedge x = x_i\} + 1}{2} \tag{3.10}$$

N	Gesamtanzahl der Sonden
$x_{1\ldots N}$	Werte aller Elemente (Sonden)
x_i	Wert eines einzelnen Elementes (Sonde) i wobei $x_i \in x_{1\ldots N}$ und $1 \leq i \leq N$
$\mathrm{card}(x)$	Kardinalität - Anzahl der Elemente mit der benannten Eigenschaft

Ein Wert x_i hat also genau eine Rangbindung b. Sie ist der Mittelwert aller Ränge r für die gilt $rgw(r) = x_i$. Beide Gleichungen können relativ einfach in MATLAB-Code umgesetzt werden.

3.3.5 Normalisierungsfunktionen

Hier werden Funktionen aufgeführt auf die im Text nicht näher eingegangen wird.

lineare Skalierung

Skalierungen sind die einfachsten Normalisierungsfunktionen. Alle Meßparameter haben einen linearen Einfluß auf den Meßwert. Nach dem Abzug der Hintergrundintensität ergibt sich folgende allgemeine lineare Skalierungsfunktion für die Vergleichsintensitäten:

$$s_n = \frac{\cdot s_p}{\kappa} \tag{3.11}$$

κ	Normalisierungsquotient
s_n	Normalisierte Signalintensität
s_p	Signalintensität, hintergrundkorrigiert

Mittelwert Die Berechnung des mittelwertbasierten Normalisierungsfaktor erfolgt durch die Addition aller hintergrundkorrigierter Signale von Gensonden und der Division dieser Summe durch die Anzahl der verwendeten Sonden auf dem Array.

$$\kappa = \bar{s} = \frac{1}{N} \sum_{i=1}^{N} s_{p,i} \tag{3.12}$$

Median

$$\kappa = \hat{s} = \begin{cases} \text{für ungerade } N\text{: } s_{r=\frac{N+1}{2}} \\ \text{für gerade } N\text{: } \frac{1}{2}\left(s_{r=\frac{N}{2}} + s_{r=\frac{N}{2}+1}\right) \end{cases} \tag{3.13}$$

Percentil/Quantil

$$0 < q < 1 : \ \kappa = \ {}^{q}\hat{s} = \ s_{r=\mathrm{int}(\frac{1}{2}+q*N)} \tag{3.14}$$

$$0\% < p < 100\% : \ \kappa = \ {}^{p}\hat{s} = \ s_{r=\mathrm{int}(\frac{1}{2}+\frac{p}{100*N})} \tag{3.15}$$

κ Normalisierungsquotient
s_r Rangwert des Quantil/Perzentil zugeordneten Ranges

lineare Regression Die lineare Regression ist keine probeninterne Skalierung. Sie setzt den Vergleich mit einem anderen Experiment/Probe voraus. Dazu wird eine Regressionsgerade durch die Signalwertepaare des Vergleichs gelegt. $S_{Exp1} \mapsto x$; $S_{Vergleichsexperiment} \mapsto y$.

$$y = \kappa * x + \gamma \tag{3.16}$$

$$\kappa = \frac{\sum_{i=1}^{N} x_i y_i - \bar{y} \sum_{i=1}^{N} x_i}{\sum_{i=1}^{N} x_i^2 - \bar{x} \sum_{i=1}^{N} x_i} \tag{3.17}$$

$$\gamma = \bar{y} - \kappa \bar{x} \tag{3.18}$$

Die Normalisierung erfolgt gegen das Vergleichsexperiment, in dem die Regressionsgleichung auf die x-Werte angewandt wird.

$$x_{norm} = \kappa * x + \gamma \tag{3.19}$$

Die normalisierten x-Werte haben jetzt die Eigenschaft gegenüber dem Vergleichsexperiment einen Regressionskoeffizienten von 1 zu haben und ein Offset von 0. Beim Vergleich mehrerer Experimente muß entweder gegen ein einzelnes Vergleichsexperiment normalisiert werden oder jedes Experiment mit jedem verglichen werden. Hier müssen die einzelnen Normalisierungskoeffizienten so angepaßt werden, daß die Standardabweichung der Regressionskoeffizienten der normalisierten Experimente möglichst klein wird.

Zentralisierung Die Zentralisierung ist eine von Zien et al. eingeführte spezielle Form der Regressionsmethode. Zur Bestimmung der Regressionsgeraden wird nur ein bestimmter Bereich der Werte verwendet. Es gehen nur Wertepaare ein, bei denen beide Werte unter einem oberen Limit (Sättigungsgrenze, Extremwerte) und über einem unteren Limit (Hintergrundeinflüsse und Rauschen) liegen. [Zien2001]

3.4 Visualisierungen

3.4.1 Scatterplot- Diagramme (SP)

Scatterplots sind zweidimensionale Diagramme. Auf der Abszisse und der Ordinate sind jeweils die Expressionssignale zweier Experimente in willkürlichen Einheiten (arbitrary units) aufgetragen. In Tabelle 3.1 sind die Formulierungen der verwendeten Spezialformen dieses Diagrammtypes aufgeführt.

3.4.2 Rang-Intensitäts-Diagramm

Das Rang-Intensitäts-Diagramm (RID) ist ein zweidimensionales Diagramm. Jede Kurve stellt ein Experiment dar. Auf der Abszisse ist der Rang eines jeden Meßwertes aufgetragen und auf der Ordinate der zugehörige nominale Meßwert selbst. (Darstellungen siehe Seite 51 Diagramm 4.25) Statt des Eigentlichen Meßwertes können auch normalisierte Intensitäten aufgetragen werden. Eine zusätzliche Variation ist das logRID bei dem logarithmierte Intensitäten aufgetragen werden.

Typ	Abk.	Abszisse	Ordinate	Lit.
$S_{i,exp1} \mapsto x$ und $S_{i,exp2} \mapsto y$				
linear	linSP	x	y	
logarithmisch	logSP	$\log_2 x$	$\log_2 y$	
Rang	rankSP	rang(x)	rang(y)	
Differenz-Summen	DiffSumSP	$(x + y)$	$(x - y)$	
MA-plot	MAP	$\log_2(\frac{x+y}{2})$	$\log_2(x - y)$	[Dudoit2000]
$S_{r,exp1} \mapsto x$ und $S_{r,exp2} \mapsto y$				
Rangintensität,lin.	rangISP	x	y	
Rangintensität,log.	ranglogISP	$\log_2 x$	$\log_2 y$	

Tabelle 3.1: Verwendete Typen von Scatterplot-diagrammen (Jeder Punkt in den 1.-5. Diagrammformen repräsentiert eine bestimmte Sonde i (Gen) mit den angegebenen Zuordnungen. In den Diagrammformen 6 und 7 sind die jeweils gleichrangigen Signale der beiden Experimente gegeneinander aufgetragen.

Kapitel 4

Ergebnisse

4.1 Hybridisierungsmodell

4.1.1 Motivation

Im folgenden Abschnitt wird aus dem vorhandenen Hybridisierungsmodell in verdünnten Lösungen [Breslauer1986] ein einfaches thermodynamisches Modell der Hybridisierungsreaktion auf dem Array entwickelt. Wie in der Einleitung beschrieben ist die Hybridisierung der zentrale Schritt der Genexpressionsanalyse. Es ist zu vermuten, daß gerade hier Parametervariationen (systematisch oder statistisch) einen starken Einfluß auf das letztendliche Signal haben. Das Modell soll helfen zu entscheiden, ob diese Einflüße kritisch für eine nachfolgende Normalisierung sind, und wie sich Variationen in diesen Parametern im Sinne einer Fehlerfortpflanzung auf die Hybridisierungssignale auswirken.

4.1.2 Das Bindungsmodell

Das Bindungsverhalten einzelner DNA-Sequenzen in verdünnten Lösungen ist sehr gut verstanden. Es können quantitative Bindungsvorhersagen gemacht werden. Es kann im Prinzip für jede Sequenz die Doppelstrangbindungskonstante K berechnet werden. Moderne Methoden benutzen dazu die „Nearest Neighbor"-Methode von Breslauer et al. [Breslauer1986], [Sugimoto1996], [SantaLucia1998]. Diese geht davon aus, daß die Bindungsenergien der gebildeten Doppelstränge im wesentlichen von den Stapelenergien (*stacking energies*) direkt benachbarter Basenpaare (*Nearest Neighbor*) abhängig ist. Weiterhin kommt noch ein Initiationsterm und ein Symmetrieterm für palindromische Sequenzen dazu (Gleichungen 4.1 und 4.2). Damit können aus jeder Sequenz die Reaktionsenthalpie $\Delta_R H$ und Reaktionsentropie $\Delta_R S$ der Reaktion berechnet werden. Mittels der bekannten Gleichungen (4.3 und 4.4) können die temperaturabhängige freie Reaktionsenergie $\Delta_R G$ und somit die Gleichgewichtskonstante K bestimmt werden.

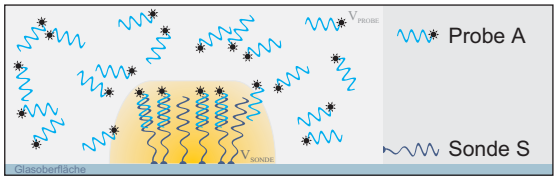
$$\begin{aligned}
\Delta H &= \Delta H_{init} + (N_{AA} + N_{TT})\Delta H_{AA} + (N_{AC} + N_{GT} + N_{CA} + N_{TG})\Delta H_{AC} \\
&\quad + \ldots + (N_{CG} + N_{GC})\Delta H_{CG} && (4.1)\\
\Delta S &= \Delta S_{init} + (N_{AA} + N_{TT})\Delta S_{AA} + (N_{AC} + N_{GT} + N_{CA} + N_{TG})\Delta S_{AC} \\
&\quad + \ldots + (N_{CG} + N_{GC})\Delta S_{CG} + \Delta S_{sym} && (4.2)\\
\Delta_R G &= \Delta_R H - T \cdot \Delta_R S && (4.3)\\
K &= e^{-\frac{\Delta_R G}{RT}} && (4.4)
\end{aligned}$$

Für die Arraymodelle werden diese Gleichungen übernommen. Im weiteren wird die Gleichgewichtssituation angenommen, und daß sich die berechneten Konstanten nur unwesentlich durch eingeführte Markierungsreste (Fluoreszenzlabel etc.) ändern. Die Verwendung dieses Modells erlaubt den Einfluß der Sequenz, der Temperatur und der Salzionenkonzentration zu simulieren. Auf dem Array wird die Situation dadurch verkompliziert, daß eine (unbekannte) Vielzahl von verschiedener Probenspezies mit einer (bekannten) Vielzahl verschiedener Sondenspezies interagiert. Obwohl die spezifische Affinität einer Sondensequenz für ihre sequenzkomplementäre Probe sehr hoch ist, können je nach experimentellen Bedingung teilkomplementäre Proben Bindungsaffinitäten zu dieser Sonde zeigen, die für nachweisbare Signale ausreichen und welche in Einzelfällen stärker sind als die der spezifischen Probe.

Parameterauswahl Als Beispiel dient die mRNA-Sequenz des Signalprotein BMP2 [GB:NM001200]. Es wurden spezifische Sonden verschiedener Länge ausgewählt (Tabelle 4.1). Die dazugeörigen Sequenzen sind im Anhang aufgeführt. Die Bindungsparameter ergeben sich nach der obigen Methode. Die verwendete Implementierung in Matlab basiert auf dem prinzipiellen Algorithmus von Breslauer [Breslauer1986] und den NN-Konstanten aus der Arbeitsgruppe von SantaLucia [SantaLucia1998], [Allawi1997], [Allawi1998A], [Allawi1998B], [Allawi1998C], [Peyret1999] (weitere Details siehe M&M). Eine zweite Reihe von 30mer Oligos mit verschiedenem GC-Gehalt wurde willkürlich ausgewählt. Weitere spezifische Parameter werden aus Tabelle 1.1 entnommen. Daraus ergeben sich Temperaturstufen von 50°C, 60°C, 68°C. Weiterhin finden solche Temperaturen Verwendung bei denen eine Änderung des Bindungsverhaltens sichtbar ist. Die Volumenauswahl reicht über 1 Größenordnung von $10mL$ bis $10\mu L$.

4.1.3 Einfache Hybridisierung

Das einfachste Modell ist die Hybridisierung einer markierten Nukleinsäurespezies (Probe) an ein immobilisiertes sequenzkomplementäres Molekül (Sonde). (Abbildung 4.1).



$$A + S \rightleftharpoons AS$$

Abbildung 4.1: Schema des einfachen Hybridisierungsmodells

Die Probe kann sich innerhalb des gesamten Volumen V_{total} bewegen. Diese verfügt über eine nachweisbare Markierung. Die Beweglichkeit der Sonde wird hingegen durch die Immobilisierung auf das von ihr erreichbare Volumen V_{sonde} beschränkt. Die Sondenmoleküle und die sondengebundenen Probenmoleküle werden nicht als feste Phase betrachtet, sondern als freibeweglich innerhalb des Sondenvolumens. Damit läßt sich das Modell einfacher entwickeln. Die nachweisbaren Molekülspezies sind die sondengebundenen Probenmoleküle (n_{AS}). Aus diesen (vereinfachten) Annahmen läßt sich Gleichung Gl. 4.5 entwickeln (siehe 3). Diese Gleichung erlaubt es die Anzahl der signalgebenden Moleküle (AS) aus den Ausgangsmengen der Sonde S und Probe A zu berechnen und so zu simulieren. Es fällt auf, daß das Sondenvolumen (V_{sonde}) nicht mehr in der Gleichung enthalten ist und somit keine Einflußgröße darstellt. Dieses liegt in der Herleitung der Gleichung begründet.

Aus der Gleichung 4.5 folgt die Abhängigkeit der Menge an sondengebundener Probe vom Hybridisierungsvolumen, den Anfangsmengen der Edukte sowie von der Gleichgewichtskonstante K. Über diese ist die Reaktion von der Temperatur, der Sequenz und der Natriumionenkonzentration abhängig (siehe

Name	Länge	GC-Gehalt	$\Delta_R H$ $\left(\frac{kJ}{mol}\right)$	$\Delta_R S$ $\left(\frac{kJ}{K \cdot mol}\right)$
$BMP2-mRNA$	1547	54%	-53704.1	-142.8
$BMP2-cDNA$	1547	54%	n.a.	n.a.
BMP2-Sonden unterschiedlicher Länge				
$B500$	500	60%	-17512.1	-46.22
$B200$	200	60%	-6939.5	-18.32
$B100$	100	60%	-3450.9	-9.12
$B50$	50	60%	-1722.5	-4.56
$B20$	20	60%	-674.5	-1.79
$B10$	10	60%	-303.6	-0.82
BMP2-Sonden für Einzel-Mismatch				
$B25_{pmCG}$	25	48%	-803.9	-2.16
$B25_{mmCA}$	25	44%	-725.2	-1.96
$B25_{mmCC}$	25	48%	-748.6	-2.03
$B25_{mmCT}$	25	44%	-747	-2.02
willkürliche Sonden mit unterschiedlichem GC-Gehalt				
$S30_{GC100\%}$	30	100%	-1118.3	-2.82
$S30_{GC80\%}$	30	80%	-1069.8	-2.77
$S30_{GC60\%}$	30	60%	-1004.9	-2.67
$S30_{GC40\%}$	30	40%	-985.2	-2.68
$S30_{GC20\%}$	30	20%	-932.9	-2.62
$S30_{GC0\%}$	30	0%	-902.7	-2.59

Tabelle 4.1: Ausgewählte Sonden für BMP-2 als Probe mit den zugehörigen Bindungskonstanten (Sequenzen im Anhang)

$$n_{AS} = \frac{1}{2}\left(n_{A0} + n_{S0} + \frac{V_{total}}{K} - \sqrt{n_{A0}^2 + 2n_{A0}\frac{V_{total}}{K} - 2n_{A0}n_{S0} + \left(\frac{V_{total}}{K}\right)^2 + 2n_{S0}\frac{V_{total}}{K} + n_{S0}^2}\right) \tag{4.5}$$

$$f = \frac{V_{total}}{K} \tag{4.6}$$

$$= \frac{V_{total}}{e^{-\frac{\Delta_R G}{RT}}} \tag{4.7}$$

$$= V_{total} e^{\frac{\Delta_R G}{RT}} \tag{4.8}$$

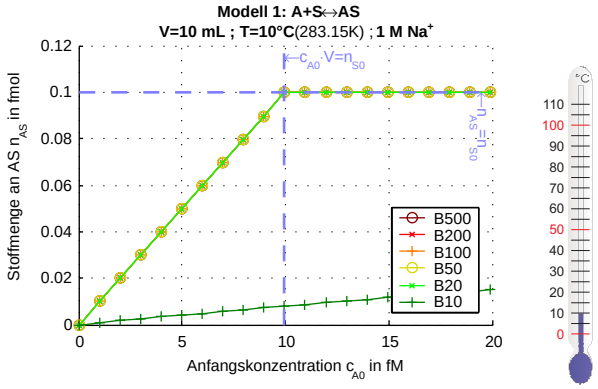


Abbildung 4.2: Einfaches Hybridisierungsmodell bei 10°C. Jede Kurve entspricht einer Sonde bestimmter Sequenzlänge (B10 - 10mer, B20 - 20mer usw.). Jeder Punkte auf der Kurve entspricht einem **virtuellem** *Experiment, in dem nur Probe und ein Sondetyp vorkommen. Die Abszisse gibt die Startkonzentration der Probe an n_{A0} an (Eingabegröße - cDNA-Konzentration von BMP2). Die Ordinate korrespondiert zur Menge an sondengebundener Probe n_{AS} (Ausgabegröße) nach Gleichgewichtseinstellung. Die waagerechte blaugestrichelte Linie gibt die Gesamtmenge der jeweiligen Sonde an (n_{S0} ist für alle Sonden gleich). Die senkrechte blaugestrichelte Linie zeigt die Konzentration an Probe an, bei der die Gesamtstoffmenge an Probe im Reaktionsvolumen gleich der Gesamtstoffmenge an Sonde ist $n_{S0} \hat{=} c_{A0} \cdot V_{total}$.Weitere Parameter V,c_{Na} sind in der Kopfzeile der Abbildung angegeben.*

22). Um die Wirkung der einzelnen Parameter zu testen, werden diese im Folgenden unter Konstanz der anderen Parameter variiert.

Temperaturabhängigkeit

Modelliert man die Temperaturabhängigkeit der Reaktion der 50mer Sonde zwischen 10°C und 95°C bei ansonsten konstanten Bedingungen zeigt sich folgendes Verhalten:

Zuerst wird das Verhalten der verschiedenen Sonden (B10-B500) und ihrer korrespondierenden Proben bei unterschiedlichen Temperaturen modelliert. Es sind jeweils $6*10^7$ Sondenmoleküle immobilisiert. Das Hybridisierungsvolumen entspricht einer Filter-hybridisierung von 10ml. Die Natriumionengesamtkonzentration ist 1 Molar. Jeder Punkt in Abbildung 4.2 entspricht einem virtuellen Experiment von einem Sonden-Proben-Paar bei einer bestimmten Temperatur und definierten Ausgangskonzentrationen. Dabei entspricht die Abszisse dem zu bestimmenden Eingangssignal (Probenkonzentration) und die Ordinate dem meßbaren Signal (Stoffmenge der sondengebundenen Probe) nach der Gleichgewichtseinstellung. Experimente gleicher Sonden-Proben-Paare sind zu einer Kurve zusammengefasst.

Bei einer Hybridisierung bei 10°C ist die Bindungskonstante so hoch, daß nahezu alle Probenmoleküle an ihren korrespondierenden Sonden binden. Alle Sonden außer der 10mer Sonde erzeugen einen übereinstimmenden linearen Anstieg des nachweisbaren Sonden-Proben-Duplex (AS) bis die molare Menge von Probe und Sonde gleich ist ($n_{A0} < n_{S0}$). Danach führt ein weiteres Erhöhen der anfänglichen Probenkonzentration (c_{A0}) zu keiner weiteren Erhöhung des Produktes AS ($n_{A0} \geq n_{S0}$). Die Produktmenge (Signal) knickt sehr scharf in die Sättigung ab. Nur bei der 10mer Sonde sehen wir einen allmählichen Übergang in die Sättigung und ein nichtlineares Verhalten für den gesamten Bereich.

Weitere Temperaturerhöhung führt zu einer Verringerung der freien Reaktionsenergie $\Delta_R G$ für alle Hybridisierungsreaktionen. Dann reichen die kurzen Sequenzen für eine effektive Duplexbindung nicht mehr aus. Bei 45°C erzeugt die 10mer Sonde kaum mehr ein Signal (Abb.4.3) und die 20mer Sonde fängt an eine allmählige Sättigungskurve zu zeigen. Deutlicher ist dieses bei weiter Temperaturerhöhung um 5°C auf 50°C (Abb.4.3). Bereits bei 60°C würde die 20mer Sonde kein Signal mehr erzeugen (Abb.4.5).

Die nächste wesentliche Änderung des Verhaltens des Sondensets ist erst bei 81°C ersichtlich. Bei dieser

Temperatur folgt die 50mer Sonde den kleineren Sonden. (Abb.4.6). Bereits 5℃ mehr läßt nur noch einen geringen Teil der korrespondierenden Probenmolküle an diese Sonde binden (Abb.4.7).

Noch extremere Temperaturen (93℃ - Abb.4.8) erzeugen auch bei der 100mer Sonde ein ähnliches Verhalten. Interessanterweise würden die 200mer und 500mer Sonden-Proben-Duplexe bei diesen Temperaturen stabil sein und nicht denaturieren. Dazu sind anscheinend noch drastischere Bedingungen (anderer Puffer, pH-Wert) notwendig.

Das Modell zeigt klar eine starke Temperatur- und Sondenabhängigkeit der Hybridisierungsfunktion. Eine wesentliche Schlußfolgerung ergibt sich aus dem Verhalten der langen Sonden. Bis 80℃ zeigen die 50mer bis 500mer Sonde ein gleiches Verhalten. Das impliziert aber auch das eine 50mer Probe an einer 500mer Sonde ein ähnliches Verhalten wie eine 100mer bis 500mer Probe komplementärer Sequenz zeigt. In diesem Fall könnte z.B. eine 200mer Probe, die eine 50% Sequenzhomologie zur Sonde hat, eine gleiches Signal hervorrufen wie die eigentliche Zielprobe. Kleinere Sonden (20mer bis 100mer) sind dafür weniger anfällig, haben aber ein Temperaturoptimum, das von Sonde zu Sonde unterschiedlich sein kann. Hier ist es von entscheidender Bedeutung, daß das Sondendesign eines Arrays nur Sonden mit gleichem Temperaturoptimum vereint.

Volumenabhängigkeit

Die meisten Arrayhybridisierungen werden bei Temperaturen von 50-60℃ vorgenommen. Variabel ist neben dem verwendeten Array auch das verwendete Hybridisierungvolumen. Dieses variiert von etwa $10mL$ bei Filtern bis $10\mu L$ bei Glasslides und Genchips. Je nach System wird daher die Gesamtmenge an eingesetzter Probe unterschiedlich stark verdünnt. Die Simulation zeigt, daß bei großen Volumina (geringen Proben-Konzentrationen) die Dissoziation begünstigt wird. Eine Probenverdünnung hat demzufolge einen ähnlichen Effekt wie eine Temperaturerhöhung (ohne Abb.).

Natriumionenkonzentration

Veränderungen in der Pufferzusammensetzung können im allgemeinen mit diesem empirischen Modell nicht berücksichtigt werden. Eine Ausnahme bildet die Natriumionenkonzentration, die bereits in den Allawi-Veröffentlichungen gut beschrieben ist. Höhere Natriumionenkonzentrationen ($1MNa^+$) haben eine stabilisierende Wirkung auf den Doppelstrang. Dabei ist der Stabilisierungseffekt sequenzunabhängig und wirkt nur auf die Phosphationen im DNA-Polyribosegerüst. Auf dieses spezielles Modell angewandt hat die Erniedrigung dieser Konzentration einen ähnlichen Effekt wie eine Temperaturerhöhung (ohne Abb.).

GC-Gehalt

In den vorherigen Beispielen hatten die Sonden zwar eine unterschiedliche Länge, aber ihr GC-Gehalt wurde konstant gehalten. Da die GC-Bindung etwa 3/2-fach stärker ist als die AT-Bindung, ist auch der Energiegewinn bei der Doppelstrangbildung bei höherem GC-Gehalt größer. Demzufolge ist das Bindungsverhalten gleichlanger Sequenzen von ihm abhängig. Höherer GC-Gehalt führt zu höheren Schmelztemperaturen und umgekehrt. (Abb. 4.9 und Abb. 4.10).

Sondenmenge

Die Menge an abgelegter Sonde variiert durch den Herstellungsprozeß des Arrays. Dabei haben die *in situ* hergestellten Oligoarrays durch ihre definierten Bindungsstellen auf dem Array und der kontrollierbaren Syntheseeffizienz einen klaren Vorteil. Bei gespotteten Arrays ist die Variation viel höher, besonders bei mechanischer Auftragung. Diese Methode hat wiederum den Vorteil, nur Moleküle gleicher Länge auf einen Spot aufzutragen. Bei Oligoarrays können Sequenzabbrüche auftreten. Es werden dadurch Sonden mit geringerer Bindungsaffinität und Selektivität erzeugt.

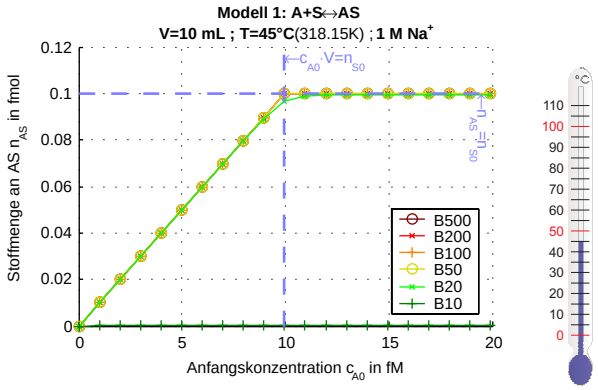


Abbildung 4.3: Einfaches Hybridisierungsmodell bei 45°C.

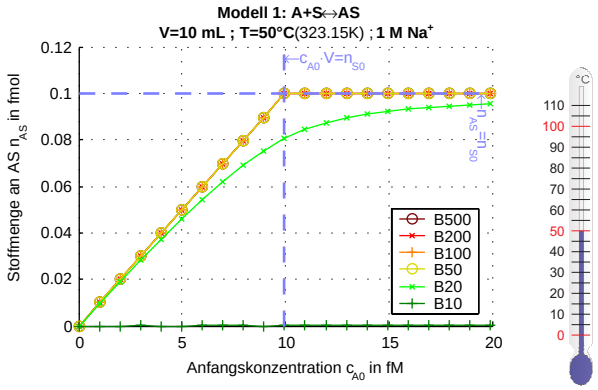


Abbildung 4.4: Einfaches Hybridisierungsmodell bei 50°C.

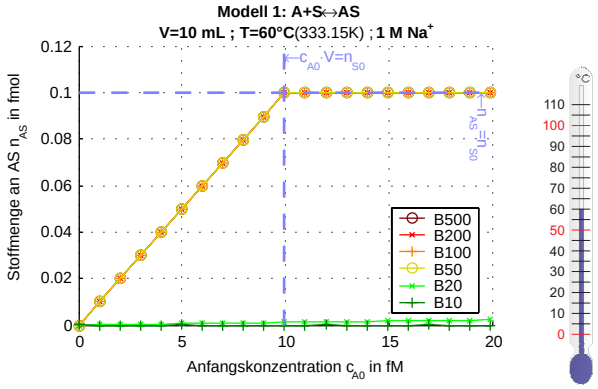


Abbildung 4.5: Einfaches Hybridisierungsmodell bei 60°C. Jede Kurve entspricht einer Sonde bestimmter Sequenzlänge (B10 - 10mer, B20 - 20mer usw.). Abszisse ⇒ Anfangskonzentration der Probe n_{A0} (Eingabegröße). Ordinate ⇒ Menge an sondengebundener Probe n_{AS} (Ausgabegröße). Siehe auch Abb.4.2.

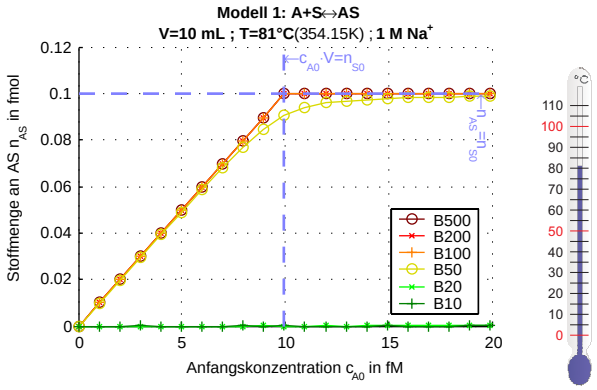


Abbildung 4.6: Einfaches Hybridisierungsmodell bei 81°C.

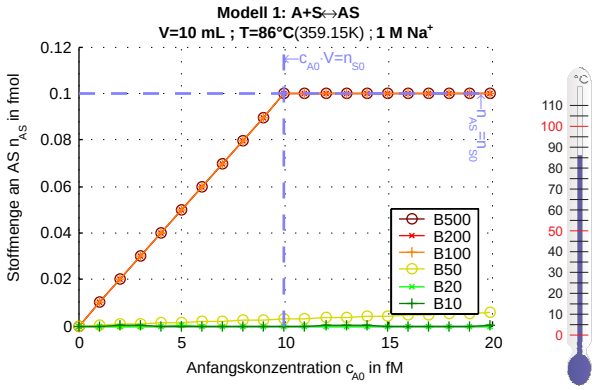


Abbildung 4.7: Einfaches Hybridisierungsmodell bei 86°C.

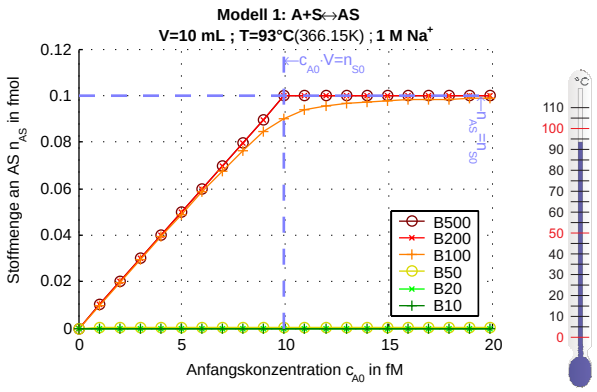


Abbildung 4.8: Einfaches Hybridisierungsmodell bei 93°C. Jede Kurve entspricht einer Sonde bestimmter Sequenzlänge (B10 - 10mer, B20 - 20mer usw.). Abszisse ⇒ Anfangskonzentration der Probe n_{A0} (Eingabegröße). Ordinate ⇒ Menge an sondengebundener Probe n_{AS} (Ausgabegröße). Siehe auch Abb.4.2.

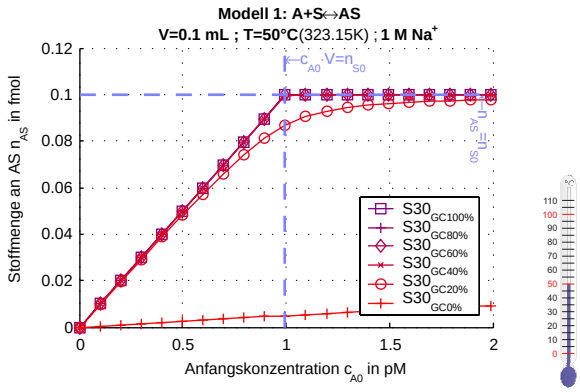


Abbildung 4.9: Einfaches Hybridisierungsmodell von Sonden/Proben Paaren mit unterschiedlichen GC-Gehalt bei 50°C.

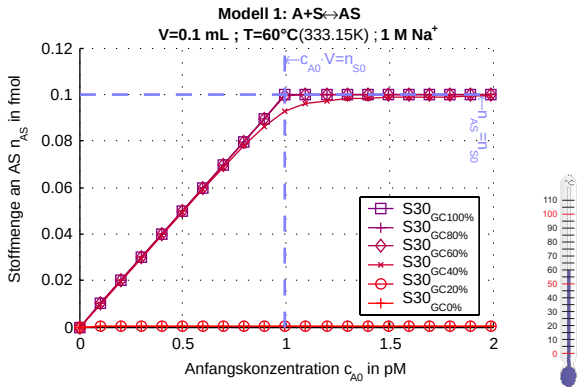


Abbildung 4.10: Einfaches Hybridisierungsmodell von Sonden/Proben Paaren mit unterschiedlichen GC-Gehalt bei 60°C. Jede Kurve entspricht einer 30mer Sonde bestimmtem GC-Gehalts. Abszisse ⇒ Anfangskonzentration der korrespondierenden Probe n_{A0} (Eingabegröße). Ordinate ⇒ Menge an sondengebundener Probe n_{AS} (Ausgabegröße). Siehe auch Abb.4.2.

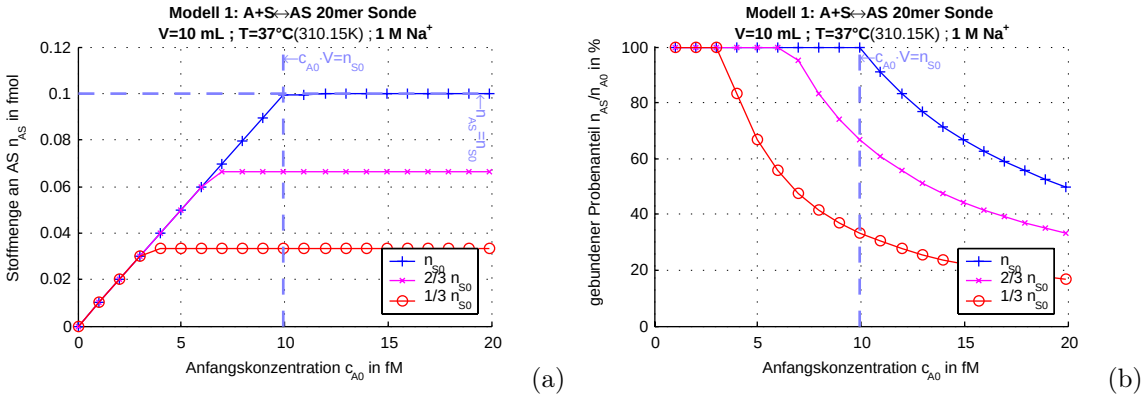


Abbildung 4.11: Einfaches Hybridisierungsmodell mit variabler Sondenmenge bei 37°C (a) Mengendiagramm siehe Abb.4.2. (b)Signalverhältnis (Anteil gebundener Probe) bei steigender Gesamtmenge an nachzuweisender Probe

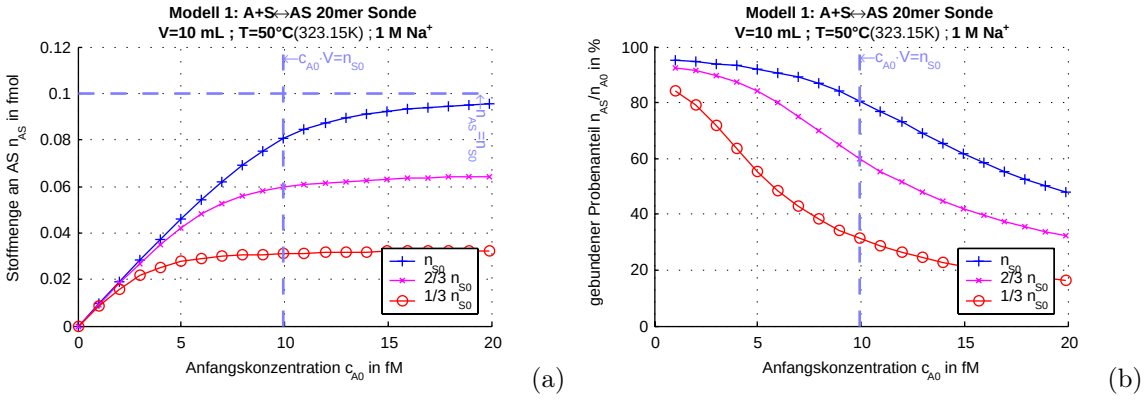


Abbildung 4.12: Einfaches Hybridisierungsmodell mit variabler Sondenmenge bei 50°C (a) Mengendiagramm siehe Abb.4.2. (b)Signalverhältnis (Anteil gebundener Probe) bei steigender Gesamtmenge an nachzuweisender Probe

Wie oben gezeigt gibt es bei niedrigen Temperaturen zwei Verhaltensabschnitte: Die Sondenmenge ist größer als die Probenmenge ($n_{A0} < n_{S0}$) und umgekehrt ($n_{A0} \geq n_{S0}$). (Niedrig heißt hier: Es treten scharfe Übergänge zwischen den Abschnitten auf). Liegen bei niedrigen Temperaturen die Probenmengen unter der niedrigsten gespotteten Sondenmengen, treten nach diesem Modell keine Schwankungen der gebunden Probenmenge (des Signals) auf. Die gesamte Probe wird gebunden. Bei höherer Probenmenge kann diese nicht mehr zuverlässig gemessen werden. Das Signal wird bei Erreichen der Sättigungsmenge ($n_{A0} = n_{S0}$) direkt abgeschnitten. Veränderungen der Probenmenge in diesem Bereich können nicht mehr von Schwankungen der abgelegten Sondenmenge unterschieden werden. Ist die Probenmenge größer als die höchste abgelegte Sondenmenge. Sind Signalunterschiede nur noch auf Schwankungen der abgelegten Sondenmenge zurückzuführen. (Abbildung 4.11)

Bei höheren Temperaturen (allmähliche Übergänge zwischen den Abschnitten) ist gar kein direkt proportionaler (linearer) Zusammenhang zwischen Gesamtprobenmenge und Signal (sondengebundene Probenmenge) ersichtlich (Abbildung 4.11.a). Im gesamten betrachteten Abschnitt ist der Kurvenverlauf von Proben- und Sondenmenge abhängig. Eine Bestimmung der zu messenden Probenmenge aus dem Signal ist hier nur noch möglich, wenn man die wirkliche Sondenmenge auf dem Array kennt und damit das Hybridisierungsverhalten simuliert. Ansonsten spiegelt das Verhältnis der Signale bei $n_{A0} \gg n_{S0}$ auch hier nur noch das Sondenmengenverhältnis wieder (Abbildung 4.12.b).

4.1.4 Grenzen des Modells (kinetische Überlegungen)

Das hier entwickelte Modell ist „nur“ ein einfaches thermodynamisches Modell. Kinetisches Verhalten kann damit nicht vorhergesagt werden. Ich gehe hier davon aus, das die Hybridisierungszeit ausreichend lang ist, so daß sich das Thermodynamische Gleichgewicht einstellen kann. In Realität gibt es eine Abhängigkeit der Signalstärke von der Hybridisierungszeit. Das liegt zum einen an der Diffusionsgeschwindigkeit der Nukleinsäuren innerhalb der Lösung und den geladenen „Molekülwolken“ um die Sondenoberflächen. Diese Diffusionsbarriere stellt im Wesentlichen eine Reaktionsbremse dar, die um so stärker wirkt, je mehr Probenmoleküle an die Sonden binden. Bei ausreichend langen Sonden ist der Energiegewinn durch die Doppelstrangbildung mit der Probe immer noch groß genug um für den Fall ($n_{A0} \ll n_{S0}$) keine Rolle zu spielen. Für $n_{A0} < n_{S0}$ stellt dieses Problem einen zusätzlichen nichtlinearen Einfluß dar, der den scharfen Übergang bei $n_{A0} = n_{S0}$ auch bei langen Sonden in einen allmählichen Übergang überführt.

Einen wesentlichen Einfluß auf die Vergleichbarkeit der Ergebnisse hat folgende kinetische Überlegung. Die hier betrachtete Hybridisierungsreaktion ist eine Reaktion 2. Ordnung, da im Idealfall zwei Moleküle (Probe und Sonde) daran beteiligt sind (Gl. 4.9). Daraus folgt die Gleichung der Produktbildungsgeschwindigkeit Gl. 4.10. Durch die Abhängigkeit dieser Geschwindigkeit von den Lösungskonzentrationen der Ausgangsstoffe ergibt sich eine zeitliche Abhängigkeit des gemessenen Hybridisierungssignal von der Sondenmenge, die durch die rein thermodynamische Betrachtung von Abschnitt 4.1.3 nicht ersichtlich ist. Wenn man also davon ausgeht, daß nur die Hinreaktion geschwindigkeitsbestimmend für das Hybridisierungssignal (Stoffmenge des Probensondenkomplex n_{AS}) ist, kann man über Gl. 4.11 und Gl. 4.12 die Gl. 4.13 ableiten. Mit dieser wäre unter Kenntnis der Geschwindigkeitskonstante $k_{hybridisierung}$ der zeitliche Verlauf der Hybridisierungsreaktion möglich.

Von einer kinetischen Kontrolle der Reaktionen kann jedoch nicht ausgegangen werden, da die Aktivierungsenergie im Wesentlichen die Energie für die Überwindung der ionischen Abstoßung zwischen den Einzelsträngen ist. Diese ist sequenzunabhängig und kann durch die Pufferbedingungen (z.B. Natriumionen-konzentration) beeinflußt werden. Entscheidend hierfür ist die lokale Anzahl der Ladungen. Diese ist damit weitgehend unabhängig von der Sequenzlänge aber abhängig von der Natriumionenkonzentration [Sabahi2001]. Ist diese Barriere überwunden reicht die Bildung weniger Wasserstoffbrücken zwischen komplementären Basen für die Initiation der Hybridisierungsreaktion (Reißverschlußmodell). ([Schwille1996]).

Bei längeren Sequenzen ($\sim$ 500 und länger) laufen die Reaktionen nicht mehr direkt ab, sondern

$$A + S \rightarrow AS \tag{4.9}$$

$$\frac{dc_{AS}}{dt} = k_{hybridisierung} \cdot c_A \cdot c_S \tag{4.10}$$

$$\frac{d\frac{n_{AS}}{V_{sonde}}}{dt} = k_{hybridisierung} \cdot \frac{c_A}{V_{total}} \cdot \frac{n_S}{V_{sonde}} \tag{4.11}$$

$$\frac{dn_{AS}}{dt} = \frac{k_{hybridisierung}}{V_{total}} \cdot n_A \cdot n_S \tag{4.12}$$

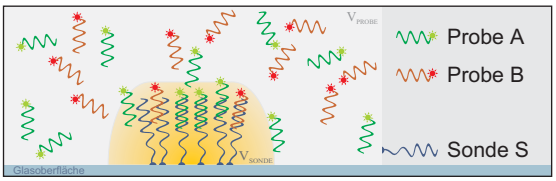
$$n_{AS} = \begin{cases} \left(\frac{n_{A0} \cdot t}{t + \frac{1}{k' n_{A0}}}\right) & ; n_{A0} = n_{S0} \\ \left(\frac{n_{A0} - n_{S0} \cdot e^{\ln(\frac{n_{A0}}{n_{S0}}) + k'(n_{A0} - n_{S0})t}}{1 - n_{S0} \cdot e^{\ln(\frac{n_{A0}}{n_{S0}}) + k'(n_{A0} - n_{S0})t}}\right) & ; n_{A0} \neq n_{S0} \end{cases} \tag{4.13}$$

$$k' = \frac{k_{hybridisierung}}{V_{total}} \tag{4.14}$$

über mehrere Zwischenstufen ([Bockelmann2002], [Causo2000], [Kafri2000], [Mukamel2002]). Dieses hat jedoch keinen entscheidenden Einfluß auf die Hybridisierung. Die Spezifität der Hybridisierung ist thermodynamischen Charakters. Die Rückreaktion ist somit der dafür entscheidende Parameter und deren Aktivierungsenergie entspricht im wesentlichen dem Energiegewinn der Hinreaktion. Wie oben gezeigt, schmelzen dadurch Doppelstränge mit mehr als 200 Basen Sequenzhomologie unter normalen Hybridisierungsbedingungen nicht mehr auf. Daraus folgt auch, daß lange Hybridisierungszeiten in diesem Sinne keinen Einfluß auf die Spezifität haben (wohl aber auf die Quantifizierung). Die hier angestellten thermodynamischen Überlegungen gelten für das System im Gleichgewicht und sind somit für die zeitliche Betrachtung Grenzwerte für lange Hybridisierungszeiten.

4.1.5 Kompetetive Hybridisierung

Das zweite Modell steht für die Hybridisierung von zwei konkurrierenden Proben (A und B) mit einer (teil-)komplementären Sonde (S). Das können zum Beispiel Nukleinsäuremoleküle einer Sequenzspezies mit unterschiedliche Markierungen, wie die unterschiedliche Fluoreszenzmarkierung bei kompetetive Mikroarrays (Abb. 4.13), sein oder Proben unterschiedlicher Sequenz, wobei eine Probe geringere Affinität zur Sonde zeigt.



$A + S \rightleftharpoons AS \;\&\; B + S \rightleftharpoons BS$

Abbildung 4.13: Schema des kompetetiven Hybridisierungsmodells I

Die Grundannahmen sind ähnlich zum ersten Modell. Die Proben A und B sind freibeweglich im totalen Reaktionsvolumen V_{total}. Die Sonde und die sondengebundenen Proben S, AS und BS sind auf das Sondenvolumen V_{sonde} beschränkt. Die Gleichgewichtskonstanten der beiden Konkurrenzreaktioen sind K_A und K_B. Zusätzlich geht man davon aus, daß die beiden Proben nicht miteinander reagieren. Die beiden meßbaren Größen sind Signale, die proportional zu den Stoffmengen der sondengebundenen

Probenmoleküle sind (n_{AS} und n_{BS}). Die eigentlich interessanten Größen sind die ursprünglich vorhandenen Stoffmengen der beiden Proben (n_{A0} und n_{B0}). Diese lassen sich über die Gleichungen Gl. 4.15 und Gl. 4.16 bestimmen. (Ableitung siehe Anhang 8.2)

$$n_{A0} = \left(\frac{V_{total}}{K_a}\frac{1}{(n_{S0}-n_{AS}-n_{BS})}+1\right)n_{AS} \tag{4.15}$$

$$n_{B0} = \left(\frac{V_{total}}{K_b}\frac{1}{(n_{S0}-n_{AS}-n_{BS})}+1\right)n_{BS} \tag{4.16}$$

Ist die Stoffmenge an Sonde sehr viel größer als die Gesamtprobenmenge ($n_{S0} \gg n_{A0}+n_{B0}$ & $n_{S0} \gg n_{AS}+n_{BS}$), dann sind die beiden Proben quasi unabhängig von einander und es gilt:

$$n_{A0} = \left(\frac{V_{total}}{K_a}\frac{1}{n_{S0}}+1\right)n_{AS} \tag{4.17}$$

$$n_{B0} = \left(\frac{V_{total}}{K_b}\frac{1}{n_{S0}}+1\right)n_{BS} \tag{4.18}$$

Ist die Stoffmenge an Sonde kleiner als oder in der gleichen Größenordnung wie die Gesamtprobenmenge ($n_{S0} = n_{A0}+n_{B0}$), dann sind die beiden Proben abhängig voneinander und von der Sondenmenge. Diese Bedingungen findet man besonders bei starker Miniaturisierung der Arrays (Glasarrays). Ausgehend von der Annahme, daß die Fluoreszenzmarkierung (z.B. A für Cy3-markierte Probe und B für Cy5- markierte Probe) keine entscheidende Veränderung der Bindungsenergie zwischen Probe und Sonde hervorruft, gilt in diesem Fall Gleichung (Gl. 4.19) und damit (Gl. 4.20). Dadurch sind die Klammerausdrücke in den Gleichungen (Gl. 4.17 und Gl. 4.18) gleich und es folgt aus Gleichung 4.21 Gleichung 4.22.

$$\Delta_R G_A = \Delta_R G_B \tag{4.19}$$

$$K_a = K_b \tag{4.20}$$

$$\frac{n_{A0}}{n_{B0}} = \frac{\left(\frac{V_{total}}{K_a}\frac{1}{(n_{S0}-n_{AS}-n_{BS})}+1\right)n_{AS}}{\left(\frac{V_{total}}{K_b}\frac{1}{(n_{S0}-n_{AS}-n_{BS})}+1\right)n_{BS}} \tag{4.21}$$

$$\frac{n_{A0}}{n_{B0}} = \frac{n_{AS}}{n_{BS}} \tag{4.22}$$

Die Gleichung 4.22 ist die wesentliche Voraussetzung für die Anwendbarkeit von kompetetiven Microarrays. Das Verhältnis zwischen der Anzahl der signalgebenden Moleküle ($\frac{n_{AS}}{n_{BS}}$) ist gleich dem Verhältnis zwischen der Anzahl der zu messenden Moleküle ($\frac{n_{A0}}{n_{B0}}$). Dieses Verhältnis ist unabhängig von der Anzahl der immobilisierten Sondenmoleküle, dem Reaktionsvolumen und anderer im 1.Modell aufgeführter Einflußgrößen. Obwohl nur schwache Rückschlüße auf die reale Menge (n_{A0}&n_{B0}) einer bestimmten Sequenzspezies möglich ist, lassen sich mit dieser Technik prinzipiell Veränderungen zwischen biologischen Proben messen ohne besondere Ansprüche an die minimale Sondenmenge und deren Reproduzierbarkeit zu stellen. Dieses ist von besonderem Vorteil, wenn sich die immobilisierte Sondenmenge schlecht bestimmen läßt. Dieses Prinzip gilt aber nur unter den gemachten Voraussetzungen. Die Modelle der folgenden Abschnitte 4.1.6 und 4.1.9 decken zusätzliche Fehlerquellen auf, durch die das gemessene Signalverhältnis stark beeinflußt werden kann.

Simulation

Geht man nicht von einer prinzipiellen Gleichheit der Reaktionsenergien aus und entwickelt die obigen Gleichungen weiter, erhält man ein Gleichungssystem, das sich über eine kubische Gleichung lösen läßt.

(Ableitung siehe Anhang 8.3) Damit lassen sich wie im 1.Modell verschiedene Parameter anwenden und testen. Eine interessante Fragestellung ist folgende: Wie verhält sich das Hybridisierungssignal n_{AS} einer Probe A (B25 perfect match) bei Zugabe einer bestimmten Menge an Probe B, wobei B eine Einzelmismatch-Variante (B25 AA-mismatch) der Probe A ist (Abbildung 4.14). Die Anfangskonzentration der Probe A ist so hoch, daß die Gesamtmenge an A, die Hälfte der Sondenmenge n_{S0} ist. Bei einer Temperatur von 50 ℃ bindet A vollständig an die Sonde. Eine kontinuierliche Erhöhung der Konkurrenzprobe B führt zuerst zu einem Anstieg des Gesamtsignals (schwarze Kurve). Wenn die Sonde vollständig gesättigt ist kann keine zusätzliche Probe binden. Der eigentlich bessere Binder A wird durch die Mismatch-Probe verdrängt.

Betrachtet man den umgekehrten Fall und legt die Mismatchprobe B ($n_{B0} = \frac{1}{2}n_{S0}$) vor, ist die Änderung des Gesamtsignals gleich. (Abbildung 4.15) Wenn alle Sondenmoleküle hybridisiert sind, kommt es zu einer Verdrängung der Mismatchprobe B durch Perfectmatchprobe A. Da A der stärkere Binder ist, wird B in stärkeren Maße und bei Erreichen von $n_{A0} \approx n_{S0}$ vollständig verdrängt. Gehören beide Proben zur gleichen Gesamtprobe (sample) sind ihre Markierungen ununterscheidbar. Nur das Gesamtsignal, das durch beide Proben hervorgerufen wird, zählt. In diesem Fall würden beide Varianten ununterscheidbare Signale liefern. Ein anderer Fall würde eintreten, würden die Proben zu zwei Gesamtproben (samples) bei einer kompetetiven Hybridisierung gehören. In einer der biologischen Proben kommt z.B. nur die Form A der Sequenz eines Genes vor und in der anderen die Form B. Hier würde das Signalverhältnis $\frac{n_{AS}}{n_{BS}}$ nicht mehr dem Verhältnis der beiden Proben $\frac{n_{A0}}{n_{B0}}$ (mRNAs) entsprechen, wenn die Sättigung der Sonde erfolgt ist ($n_{A0} + n_{B0} > n_{S0}$. Die Gleichung Gl.4.22 gilt dann nicht mehr (Abbildung 4.14.b und 4.15.b).

4.1.6 Kreuzhybridisierungen an alternativen Sonden

Das dritte Modell ist die Umkehrung des zweiten Modells, wobei eine Probe A mit zwei Sonden hybridisieren kann. Die beiden Sondenarten unterscheiden sich im Bindungsverhalten. Die Sonde S_{pm} ist hier der stärkere Binder für Probe A. Die alternative Sonde Smm hat eine geringere Affinität zur Sonde. Prinzipiell könnten S_{mm} und S_{pm} auch für irgendeine Sonde stehen, die irgendeine Bindungsaffinität zur Probe haben. Diese, wie auch die Reaktionsprodukte AS_{mm} und AS_{pm}, sind auf die jeweiligen Sondenvolumina beschränkt. Die Probe kann sich im Gesamtvolumen bewegen.

Dieses Modell ist besonders interessant, weil es Kreuzhybridisierungen einer Probe zu unterschiedlichen Sonden beschreibt und somit Aussagen über die Spezifität der Signale in Arrayhybridisierungen treffen kann. Als spezielles Beispiel dient wieder eine 25mer Probe A (B_{25}) die mit zwei 25mer Sonden (B_{25pmAT} und B_{25mmAA}) hybridisieren kann (Abb. 4.16). Der Unterschied zwischen den Sonden ist somit nur eine Base. Die Sondenmenge n_{SO} ist für beide in diesem Falle gleich. Bei schrittweiser Erhöhung der Anfangsprobenkonzentration bleibt die Probe bevorzugt an der Perfectmatch-Sonde haften (T=50℃ - Abb. 4.17.a).

Erst wenn diese gesättigt ist ($n_{A0} \approx n_{S_{pm}0}$),erfolgt eine Bindung an die Mismatch-Sonde. Daraus folgt, daß im Falle $n_{A0} > n_{S_{pm}0}$ es zu Fehlsignalen (Kreuzhybridisierungen) kommt, wenn eine andere alternative Sonde auf dem Array eine gewisse Affinität zur Probe hat (Abb. 4.17.a). Dieses Fehlsignal ist additiv zu dem Signal der alternativen Sonde (siehe Modell 2). Die konkreten Parameter beinhalten bei dem gezeigten Diagramm halb so viel mismatch- wie perfectmatch Probe, so daß auch noch die Sättigung der mismatch Sonde bei $n_{A0} \approx (n_{S_{pm}0} + n_{S_{mm}0})$ gezeigt ist. Bei höherer Temperatur (T=68℃ - Abb. 4.17.b) bindet zwar immer noch die perfect-match-Probe, aber mit zu geringer Affinität. Die Kreuzhybridisierung im Fall $n_{A0} > n_{S_{pm}0}$ ist durch Temperaturerhöhung nur auf Kosten des erwünschten perfect-match-Signals zu verhindern.

Mehrfach gespottete Sonden Eine andere Anwendung des Modells ist folgende einfache Fragestellung: Wie verhält sich die Probe bei Vorhandensein von zwei gleichen Sonden (z.B. bei mehrfach gespotteten Arrays)? Beide Sonden unterscheiden sich nur in der Menge $n_{S_X0} = \frac{1}{2}n_{S_Y0}$. In Abbildung

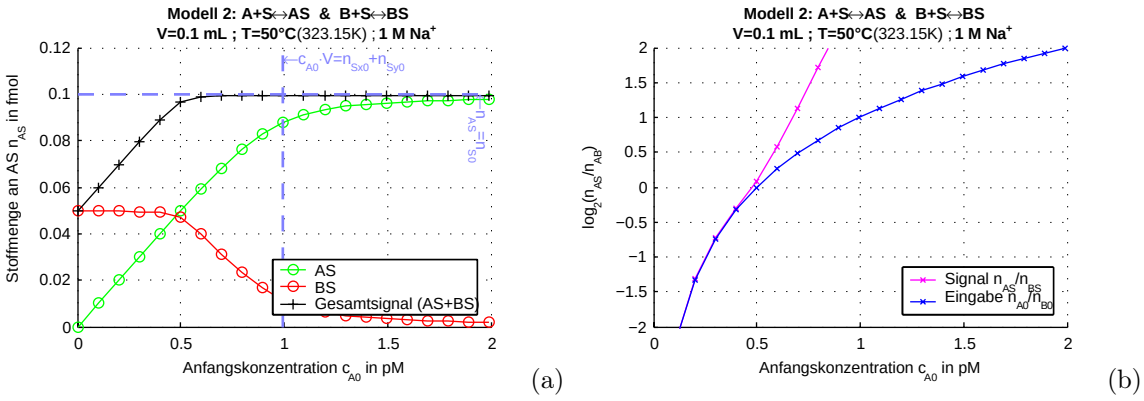


Abbildung 4.14: Kompetetives Hybridisierungsmodell mit vorgelegter Perfectmatchprobe A und mit variabler Menge der Mismatchkonkurrenzprobe B bei 50°C (a) Mengendiagramm siehe Abb.4.2. (b) Signalverhältnis zwischen A und B (blau ↦ Gesamtmengenverhältnis/ Eingangsgröße; magenta ↦ Verhältnis der gebundenen Proben/ Ausgangsgröße)

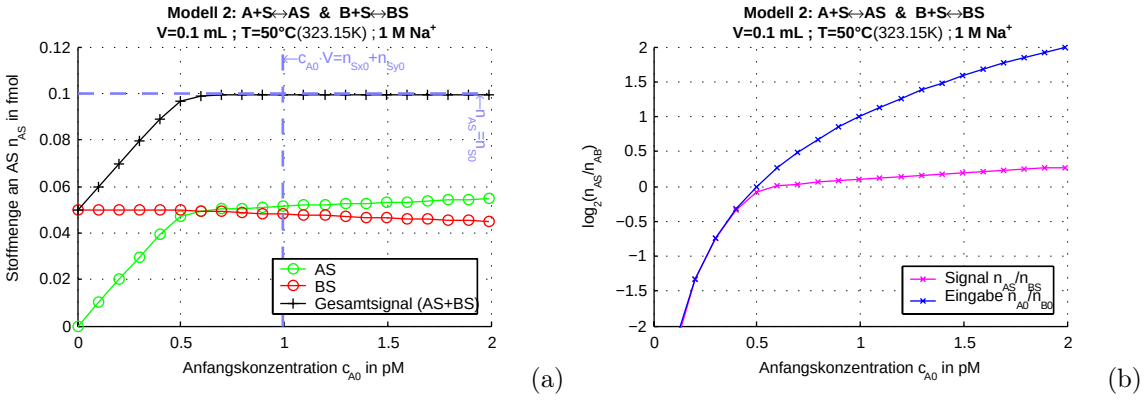
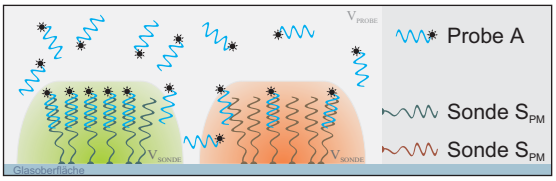


Abbildung 4.15: Kompetetives Hybridisierungsmodell mit vorgelegter Mismatchkonkurrenzprobe B und mit variabler Menge der Perfectmatchprobe A bei 50°C (a) Mengendiagramm siehe Abb.4.2. (b) Signalverhältnis zwischen A und B (blau ↦ Gesamtmengenverhältnis/ Eingangsgröße; magenta ↦ Verhältnis der gebundenen Proben/ Ausgangsgröße)



$$A + S_{pm} \rightleftharpoons AS_{pm} \;\&\; A + S_{mm} \rightleftharpoons AS_{mm}$$

Abbildung 4.16: Schema des kompetetiven Hybridisierungsmodells II

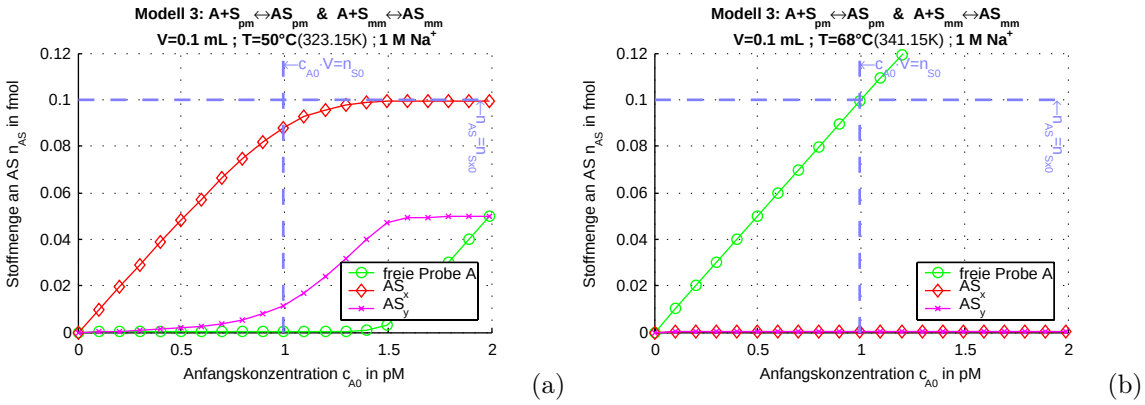


Abbildung 4.17: Kompetetives Hybridisierungsmodell mit einer Probe und zwei alternativen Bindungsmöglichkeiten unterschiedlicher Bindungsaffinität (25mer mm- und pm-Sonden) bei (a)50°C (b)68°C (Diagrammprinzip siehe Abb. 4.2)

4.18.a sieht man, daß das Hybridisierungssignal bei beiden Sonden bei Erhöhung der Probenkonzentration in gleichem Verhältnis zur jeweiligen Sondengesamtmenge steigt. Beide Sonden werden „gleichzeitig“ gesättigt bei $n_{A0} = n_{SX0} + n_{SY0}$. Das dieses Verhältnis über den gesamten Konzentrationsbereich der Probe und bei verschieden Temperaturen gleichbleibt zeigt Abbildung 4.18.c, auch wenn sich das generelle Hybridisierungsverhalten ändert (Abbildung 4.18.b).

Aus diesen Überlegungen folgt: Abweichungen von Signalen mehrfach gespotteter Sonden reflektieren nur Variationen im Spotprozeß. Sie sagen nichts über die Variabilität der Proben aus.

4.1.7 Waschprozesse

Eine weitere interessante Fragestellung, die mit diesen Modellen untersucht werden kann, ist das Verhalten der hybridisierten Proben beim nachfolgenden Waschprozeß. Hier werden hauptsächlich ungebundene Proben entfernt, die unspezifische Signale beim Scannen des Arrays hervorrufen würden. Aber auch kreuzhybridisierte Proben sollen durch das Waschen entfernt werden. Der Waschprozess wird dadurch simuliert, daß die Anzahl, der bei der Hybridisierung gebundenen Probenmoleküle, als Probengesamtmenge für eine erneute Hybridsierungssimulation mit veränderten Parameter verwendet wird.

Die im ersten Schritt nicht hybridisierten Probenmoleküle wurden weggewaschen und sind für den Waschvorgang nicht mehr vorhanden. Die erneute Hybridisierung (Waschschritt) erfolgt bei gleicher Temperatur aber größeren Reaktionsvolumen (1000faches Volumen - entspricht $100mL$ bei einer ursprünglichen $100\mu L$ Hybridisierung). In der Simulation wird die Zielprobe A als komplementäres 50mer der Sonde definiert (B50). Als Kreuzhybridisierungsprobe B wird das 20mer (B20) eingesetzt (oder z.B. ein 50mer mit 40% komplemetärer Sequenz). Dieses ist in der 1.Hybridisierung immer mit $n_{B0} = n_{S0}$

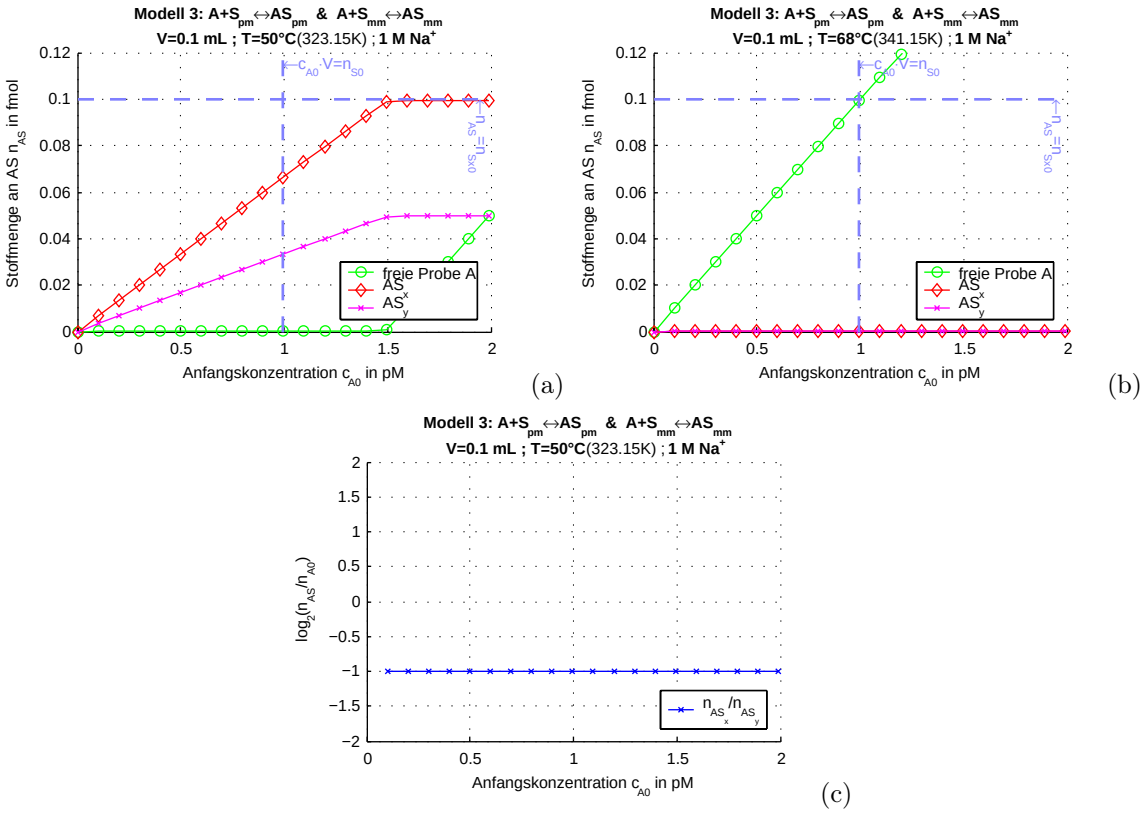


Abbildung 4.18: Kompetetives Hybridisierungsmodell mit einer Probe und zwei alternativen Bindungsmöglichkeiten gleicher Bindungsaffinität (25mer mm- und pm-Sonden) bei (a)50°C (b)68°C (Diagrammprinzip siehe Abb. 4.2.) (c)Duallogaithmisches Verhältnis zwischen den Signalen beider Spots

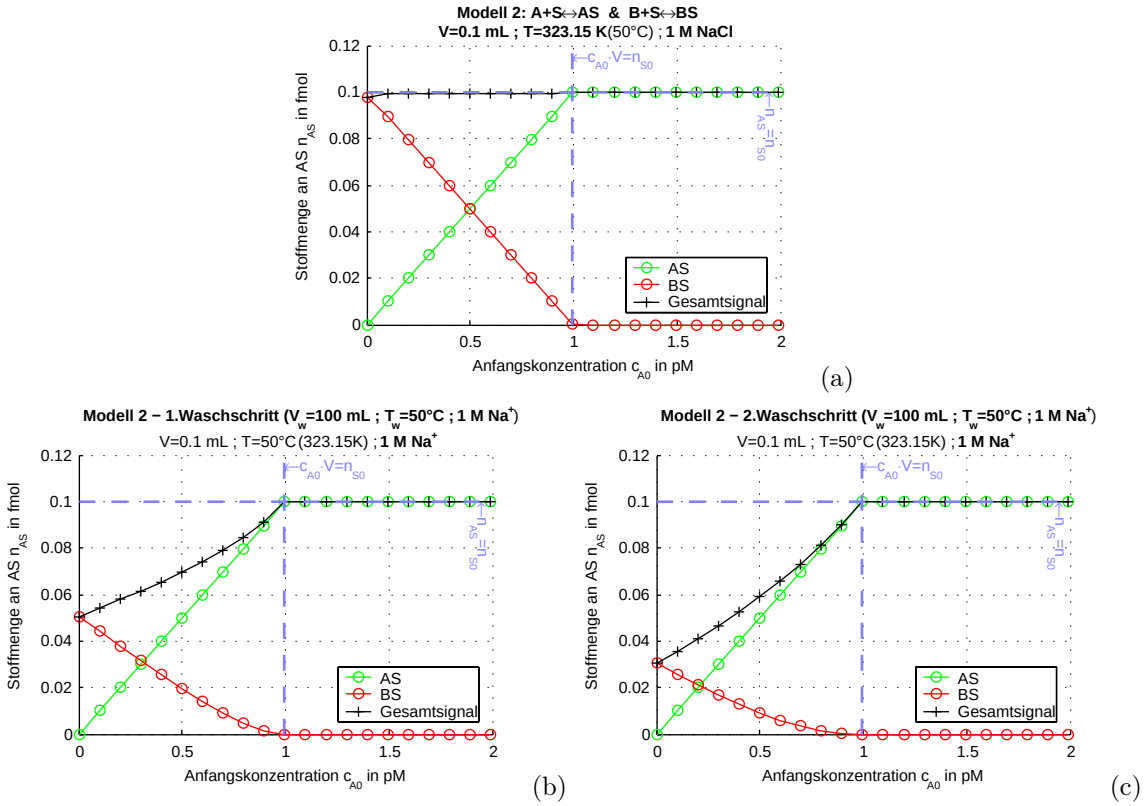


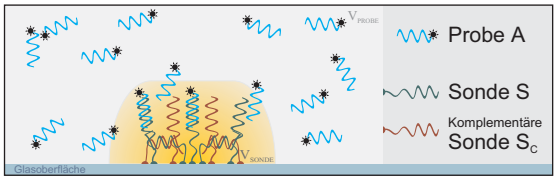
Abbildung 4.19: Kompetetives Hybridisierungsmodell mit einer Sonde und zwei möglichen Bindern A und B. B ist die kreuzhybridisierende Probe und ist bei steigender Konzentration von A konstant in der 1.Hybridisierung vorhanden (a) Hybridisierung (b) 1.Waschschritt (c) 2.Waschschritt (Diagrammprinzip siehe Abb. 4.2)

vorhanden. Wie man in Abbildung 4.19(a) bis Abbildung 4.19(c) wird durch das Waschen vorwiegend die Probe an der Mismatch-Sonde entfernt.

4.1.8 Doppelsträngige Sonden

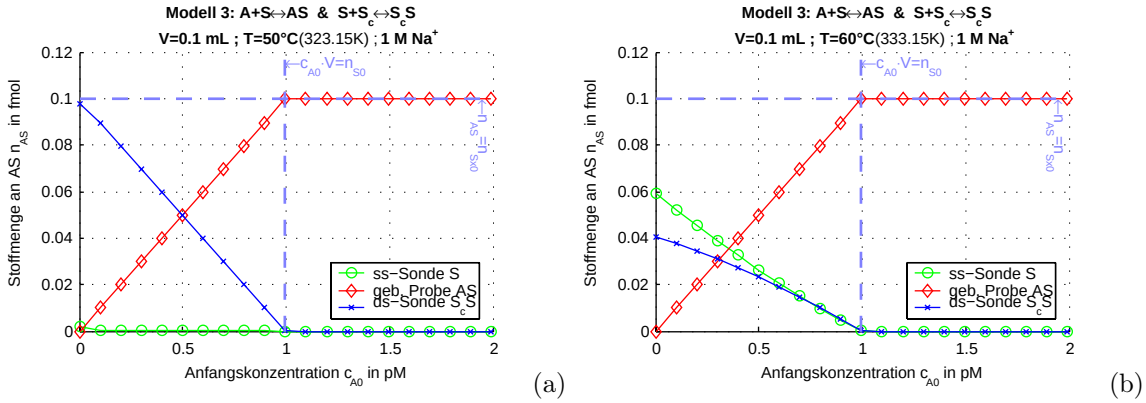
Nicht immer kann man von einzelsträngigen Sonden auf dem Array ausgehen. Oft werden doppelsträngige pcr-Produkte zur Immobilisierung verwendet. Sind die molekularen Ankergruppen für die Oberflächenbindung (z.B. endständige Aminogruppen am Polyetherlinker des Primers) an beiden Strängen vorhanden, werden beide Stränge immobilisiert. Das kann den Vorteil haben, daß nun für cDNA und mRNA gleichermaßen komplementäre Sonden vorhanden sind und damit diverse Probenamplifikationsmethoden verwendet werden können. Der große Nachteil ist jedoch die nun mögliche Konkurrenzreaktion zwischen diesen beiden Strängen. (illustriert in Abb. 4.20).

Wird diese Reaktion nicht durch zusätzliche Arraybehandlung verhindert (z.B. UV-Crosslinking an den Träger), führt sie zu nichtlinearen Signalverhalten der Hybridisierungsreaktion. Die Simulation dieses Verhaltens ist in (Abbildung 4.21.a) dargestellt. Der eigentliche probenkomplementäre Sondestrang S kann nun entweder mit der einzelsträngigen Probe A hybridisieren oder mit seinem coimmobilisierten Komplementärstrang S_c. Erweiterte Modellanahmen zu Modell 1 sind: Die beiden Sondenstränge (S und S_c) und das doppelsträngige Sondenmolekül S_cS sind in ihrer Bewegung auf das Sondenvolumen



$$A + S \rightleftharpoons AS_{pm} \;\&\; S_c + S \rightleftharpoons S_c S$$

Abbildung 4.20: Schema der Hybridisierung mit Doppelstrangsonden



(a) (b)

Abbildung 4.21: Kompetetives Hybridisierungsmodell mit einer Probe A und einer Sonde, bei der beide Stränge S und S_c vorhanden sind und miteinander hybridisieren können. bei (a)50°C (b)60°C (Diagrammprinzip siehe Abb. 4.2)

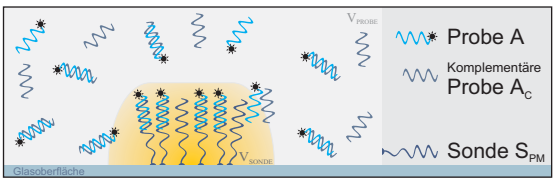
beschränkt. Im Gegensatz zu den anderen Modellen bleibt das Sondenvolumen als zusätzlicher Parameter erhalten. Dieser Parameter ist schwer abzuschätzen, daher wird von einem willkürlichen Wert von 1/1000 des Gesamtvolumens ausgegangen. Es wird weiterhin angenommen, daß der Doppelstrangbereich der Sonde gleichlang der des Sonden-Proben-Hybrids ist. Eine Erhöhung der Temperatur beeinflußt beide Reaktionen. Zwar liegt bei 60°C mehr einzelsträngige Sonde vor, aber das Gleichgewicht der Hybridisierungreaktion mit der Probe wird zur Dissoziation hin verschoben (Abbildung 4.21.b).

Ist der hybridisierbare Doppelstrangbereich der Sonde viel kleiner als die des Sonden-Proben-Hybrids entspricht das Diagramm dem Verhalten der Signalbildung n_{AS} dem der Einzelstrangsonde. (keine Abbildung)

4.1.9 Doppelsträngige Proben

Eine ähnliche Beeinflussung der Hybridisierung wie im vorhergehenden Fall, hat die Zugabe einer doppelsträngigen Probe. Diese kann z.B. durch einen Amplifizierungsschritt entstanden sein und es erfolgte keine anschließende Strangtrennung. Dadurch wird wiederum eine zusätzliche Konkurrenzreaktion eingeführt (Abb. 4.22).

Die Simulation selbst basiert auf dem Grundmodell. Die Probe A, der Komplementärstrang A_c und das konkurrierende Hybrdisierungsprodukt A_cA können sich im Gesamtvolumen verteilen. Durch die alternative Reaktionsmöglichkeit mit dem Komplementärstrang A_c wird auch bei diesem Modell ein nicht lineares Signalverhalten erzeugt. (Abbildung 4.23)



$$A + S \rightleftharpoons AS_{pm} \;\&\; A_c + A \rightleftharpoons A_c A$$

Abbildung 4.22: Schema der Hybridisierung mit Doppelstrangproben

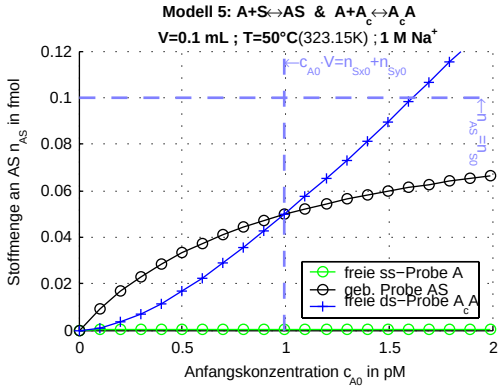


Abbildung 4.23: Kompetetives Hybridisierungsmodell mit einer Sonde S und einer Probe, bei der beide Stränge A und A_c vorhanden sind und miteinander hybridisieren können (Diagrammprinzip siehe Abb. 4.2)

4.1.10 Schlußfolgerungen und Fehlerabschätzung

Die Hybridisierung hat einen starken systematischen Einfluß auf die Ergebnisse. Obwohl allgemeine Kriterien abgeleitet werden können, ist eine quantitative Erfassung aller relevanten Einflüsse nahezu unmöglich. Auch bleibt das letztendliche Verhalten der Nukleinsäuren auf dem Chip ohne weitergehende Experimente und theoretische Behandlung unbestimmt. Daher ist die Hybridisierungsfunktion nicht explizit zu bestimmen.

Trotzdem konnte gezeigt werden, daß unter bestimmten Bedingungen durchaus von einem linearen Hybridisierungssignal ausgegangen werden kann. Die Bestimmung der Parameter unter denen diese Linearität gilt, ist im Wesentlichen die Aufgabe des Experimentators und des Plattformherstellers. Werden die Hybridisierungen unter nicht optimalen Bedingungen ausgeführt, ist nicht einmal mehr von einer Eindeutigkeit des Hybridisierungssignals auszugehen.

Das Hybridisierungssignal einer Sonde ist die Gesamtheit aller an dieser Sonde gebundenen signalgebenden Proben. Die Hybridisierungsfunktion ist abhängig von der Konzentration der korrespondierenden Probe, der Menge der Sondenmoleküle dieser Sonde, aber auch von der Konzentration aller anderen Proben und den Mengen aller anderen Sonden, dem Hybridisierungsvolumen, Hybridisierungszeit und den jeweiligen Gleichgewichtskonstanten aller möglichen Hybridisierungen zwischen allen vorhandenen Proben und allen vorhandenen Sonden.

Die Fehleranfälligkeit der letztendlich gewonnenen Daten durch das Hybridisierungssystem liegt besonders in den systematischen Fehlern, die durch Bedingungen außerhalb des Optimums eingebracht werden. Diese Fehlerart ist daher auch mit statistischen Mitteln schlecht zu behandeln. Sie kann die eigentlich biologischen Unterschiede so stark überlagern, das keine Rekonstruktion dieser Unterschiede aus den Daten möglich ist. Aber auch innerhalb optimaler Parameter sind lokale Fehlsignale aufgrund

von Kreuzhybridisierungen möglich.

4.2 Visualisierung und Bewertungskriterien

4.2.1 Motivation

Visualisierungen sind für die Bewertung von komplexen Daten von grundlegender Bedeutung. Durch geeignete Diagramme werden Zusammenhänge deutlicher und bestimmte Eigenschaften der verwendeten Daten werden herausgestellt. Dabei ist die Suche nach der „besten“ Darstellungsform ein Erkundungsprozess.

4.2.2 Scatterplot- Diagramme

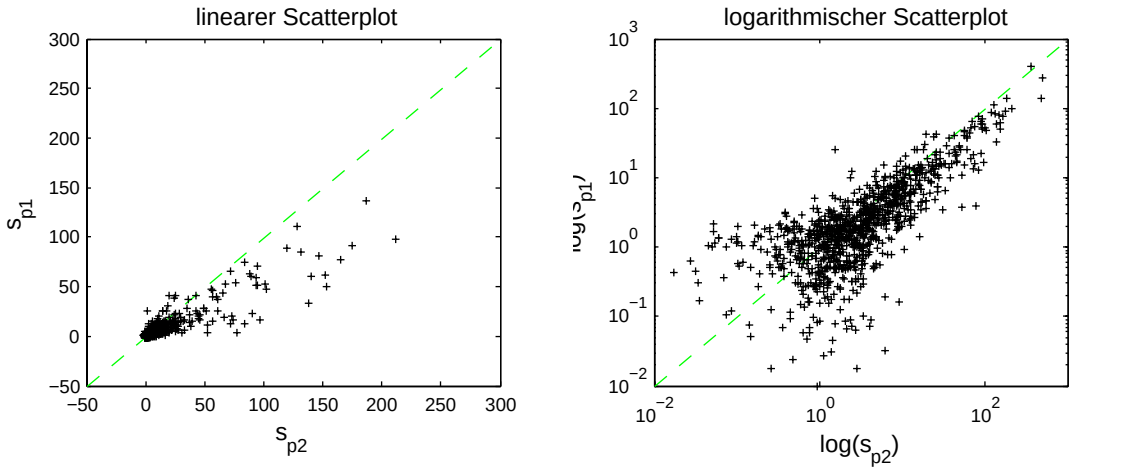


Abbildung 4.24: Beispiel eines ***(a) linearen Scatterplots*** *und* ***(b) logarithmischen Scatterplots*** *zweier HAA1.2 Hybridisierungen zweier Lungenproben. Die Identitätsgerade ist grün dargestellt*

Die einfachste Darstellung der GEA-Daten ist ein Scatterplot (SP) bei dem die Intensitäten zweier Experimente linear gegeneinander aufgetragen werden. Als Beispiel dient Abbildung 4.24.a. Hier sind die hintergrundkorrigierten Signale zweier Leukämieproben (M&M) miteinander verglichen.

Man sieht eine relative gute Korrelation der beiden Daten im unteren Bereich. Es fällt auf, daß das ein Experiment eine generell stärkere Intensität zeigt. Dieser generelle Unterschied bei sonst so ähnlichen Proben ist auf unterschiedliche Meßparameter zurückzuführen (siehe Kapitel 2). Wenden wir zum Beispiel eine Mittelwertnormierung an wird die Lage der Scatterwolke so verändert, daß die Mehrzahl der Werte sich um die Identitätsgerade ($y = x$) gruppieren. Auf Grund des starken Größenunterschiedes der Werte eines Experimentes läßt sich dieses Verhalten noch besser durch eine logarithmische Darstellung (logSP - Abbildung 4.24.b) zeigen.

Aus diesen Darstellungen lassen sich folgende Kriterien ablesen: Wie gut korrelieren die Daten? Welche Datenpunkte liegen außerhalb der Korrelationsbereiches (Interessante Gene)? Wie nah liegt die Hauptmenge der Werte an der Identitätsgeraden? Diese Kriterien lassen sich gut in mathematische Formulierungen umschreiben (z.B. Korrellationskoeffizient). Doch gerade das dritte Kriterium ist mit dieser Darstellung nicht so einfach zu evaluieren. Dazu ist es besser eine Transformation vorzunehmen und die

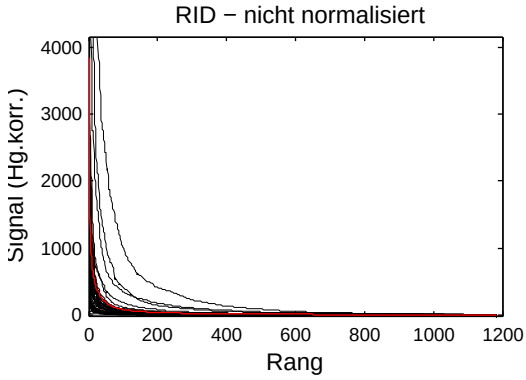


Abbildung 4.25: Rang-vs.Intensität-Kurven verschiedener Proben (RID). Die rote Kurve zeigt den Mittelwert der gleichrangiger nicht normalisierten Intensitäten. (Datenquelle: HAA1.2 Membranen mit Proben von Lunge, Niere und Blut 3)

Scatterplots um $-45°$ zu drehen. Damit sind die Veränderungen bezüglich der Identitätsgeraden auf der Ordinate aufgetragen und die Summe der Expressionstärke auf der Abszisse. Mit linearer Skalierung ist es das Summen-Differenzen-Diagramm. Mit logarithmischer Skalierung würden die logaritmierten relativen Veränderungen versus der Summe der logarithmierten Expressionswerte aufgetragen werden. Diese logarithmische Darstellung ist auch als MA-Plot in der GEA sehr gebräuchlich. Ein kleiner Unterschied existiert dabei: auf der Abszisse wird, anstelle der Summe, der Mittelwert der Logwerte aufgetragen [Dudoit2000], [Yang2002A].

4.2.3 Verteilungsdiagramme

Vergleicht man mehrere Experimente miteinander, werden Scatterplots sehr schnell unübersichtlich. Für die Normalisierung sind globale Kriterien wichtiger als die Feinstruktur im Scatterplot. So liefern zum Beispiel Rang-Intensitäts-Diagramme eine Überblick über die globale Werteverteilung. Dazu werden die Werte nach der Hintergrundkorrektur nach absteigender Intensität geordnet. Jeder Intensität wird je nach Position eine Rangzahl zugeordnet. Die höchste Intensität hat den Rang 1, die zweithöchste Rang 2, usw... Die niedrigste Intensität hat den Rang N, wobei N die Gesamtzahl aller betrachteten Gene ist. Die resultierenden Rang- Intensitäts- Kurven sind in Abbildung 4.25 dargestellt. Jede Kurve repräsentiert ein Arrayexperiment.

Für den Vergleich der Normalisierungsmethoden habe ich mit diversen Scatterplots begonnen, bei denen immer zwei Experimente miteinander verglichen werden. Dabei war es schwierig mehrere Experimente miteinander vergleichbar zu machen. Während dieser Erkundungsphase stieß ich auf auf eine nützliche rangbasierte Darstellungsmethode: das Rang-Intensitäts-Diagramm. Dieses und die aus ihm ableitbaren Kriterien sollen im folgenden Abschnitt eingeführt werden.

Es fällt auf, daß alle Genexpressionsdaten eines Arrays ähnliche Rangkurven zeigen. Die hohen Variationen im Gesamtsignal werden durch unterschiedliche Parameter hervorgerufen. Diese Ähnlichkeit wird daher offensichtlich, wenn man eine Normalisierung anwendet. Wie oben kommt dazu die Mittelwertnormierung zum Einsatz (Abb. 4.26.a). Für alle hier betrachteten Datensätze gilt folgende Verteilung: Es gibt sehr viele gering- oder nicht exprimierte Gene (genauer signalliefernde gebundene Probenmoleküle), einen mittelstark exprimierten Bereich und sehr wenige hochexprimierte Gene. Jede Kurve hat eine negativ exponentielle Form mit einer zusätzlichen Krümmung ins Negative bei großen Rangzahlen.

Unter der Annahme, daß diese Ähnlichkeit eine allgemeine Eigenschaft ist, können Rang-Intensitäts-

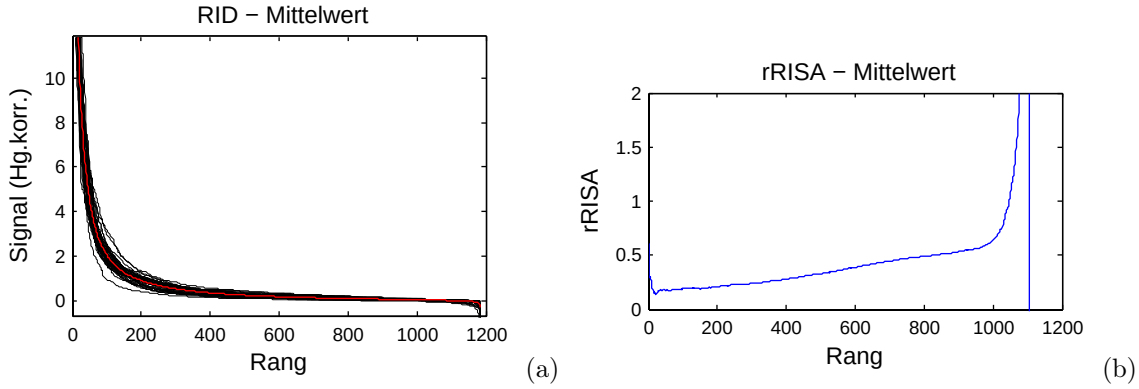


Abbildung 4.26: ***(a) RID:*** *Rang-Intensitäts-Kurven mit mittelwertnormalisierten Daten Die rote Kurve zeigt den Mittelwert der gleichrangiger Intensitäten.* ***(b) rRISA:****relative Rangintensitäts-Standardabweichung von mittelwertnormalisierten Daten (Datenquelle: HAA1.2 Membranen mit Proben von Lunge, Niere und Blut)*

Diagramme als Evaluierungswerkzeug für Normalisierungen benutzt werden. [Kroll2002B]

Obwohl von diesen Darstellungen keine Schlußfolgerung für ein einzelnes Gen gemacht werden kann, gibt das Rang- Intensitäts- Diagramm einen guten Eindruck über allgemeine Eigenschaften der einzelnen Experimente. Das Rauschen der Daten ist geglättet, da das Rauschen im wesentlichen den Rang eines Genes lokal verändert, aber weniger Einfluß auf die Intensität benachbarter Ränge hat. Für diese globale Ansicht spielt es keine Rolle, welches Gen welchen Rang besetzt. Das Meßrauschen hat einen absoluten Einfluß auf alle Werte, jedoch besonders auf niedrig- oder nicht exprimierte Werte. Dieses zeigt sich durch die negativen Werte, welche bei allen Experimenten nach der Hintergrundkorrektur auftreten.

Die Rang-Intensitäts-Kurven sind eng verwandt mit den Verteilungsfunktionen der einzelnen Experimente. Sie sind zueinander x-y-gespiegelt. Wobei bei der Verteilungsfunktion der normierte Rang als Ordinate verwendet wird. (Der normierte Rang ist der jeweilige Rang, dividiert durch die Gesamtanzahl der Messungen.) Für den Experimentvergleich, wie er hier vorgenommen wird, ist das RID die leichter interpretierbare Form.

Aus diesem Diagramm lassen sich Kriterien für die Normalisierung ableiten: Die Standardabweichung um den Mittelwert der gleichrangigen Intensitäten (RISA → **R**ang- **I**ntensitäts- **S**tandard**a**bweichung), siehe Gleichung 4.23, und die zugehörige relative Standardabweichung (rRISA → **r**elative **RISA**), siehe Gleichung 4.24. Beide Werte sind ein Maß für die lokale Abweichung der Rang-Intensitäts-Kurven von einander. Unter der Annahme der Gleichheit der Verteilungsfunktionen ist dieses ein Maß des lokalen Normalisierungsfehlers.

$$\text{RISA}: \quad \sigma_r = \sqrt{\frac{\sum_{i=1}^{n_e} \left(I_{i,r} - \bar{I}_r\right)}{n_e - 1}} \tag{4.23}$$

$$\text{rRISA}: \quad \upsilon_r = \frac{\sigma_r}{\bar{I}_r} \tag{4.24}$$

Beide Kriterien lassen sich auch einzeln in einem Diagramm darstellen. Da gerade relative Veränderungen in den Daten interessant sind, ist das rRISA-Diagramm die bessere Wahl. Der in Abbildung 4.26.b gezeigte Vergleich unterstreicht die Nützlichkeit dieses Diagramms beim Vergleich von Normalisierungsmethoden.

4.3 Abschätzung des additiven Rauschens des Detektionssystems

4.3.1 Motivation

Es fällt auf, daß Arrays, die ein schwaches Gesamtsignal haben, meist die größte Anzahl negativer Werte besitzen. (Schwache Signale sind dadurch gekennzeichnet, daß sie sich visuell schlecht vom Hintergrundrauschen absetzen.) Der Grund für das Auftreten negativer Signale liegt in der Subtraktion des geschätzten Hintergrundsignals vom Spotsignal (siehe Systemanalyse). In dem Falle, in dem zufällig (Rauschen) oder systematisch (Sockelintensitäten) das Hintergrundsignal größer ist als das Spotsignal, ist das resultierende Signal negativ.

Im Folgenden wird angenommen, daß je größer der absolute Rauschfehler ist, desto größer ist der Anteil negativer Werte am Gesamtsignal.(Gl. 4.25).

$$\sigma_{abs} \quad \propto \quad \left| \frac{\sum x_-}{n_{x-}} \right| \tag{4.25}$$

(Im folgenden wird $\left| \frac{\sum x_-}{n_{x-}} \right|$, der absolute Mittelwert der negativen Werte, als $\bar{x}_-$ bezeichnet. Entsprechend ist der Mittelwert der positiven Werte $\bar{x}_+$)

4.3.2 Testdaten

Als Modell für die rauschfreie empirische Verteilungsfunktion der Arraydaten wird eine einfache exponentielle Funktion benutzt (M&M S. 23). Dabei wird davon ausgegangen, daß die negativen Werte nicht in der ursprünglichen Verteilungsfunktion enthalten sind. Der Rauscheinfluß wird durch Addition einer normalverteilten Rauschfunktion simuliert. In Bild 4.27.a sind die daraus resultierenden Rang-Intensitäts-Kurven dargestellt. Jede Kurve repräsentiert die exponentielle Verteilungsfunktion plus einem Rauschen mit definierter Standardabweichung σ_{abs}.

Laut der Vermutung müßte sich eine Korrelation zwischen dem Mittelwert der negativen Werte jeder einzelnen Kurve und der eingegangen Standardabweichung des Rauschens feststellen lassen. Dieses ist der Fall, wie Abbildung 4.27.b belegt.

$$\sigma_{abs} \quad = \quad f_{exp} \bar{x}_- \tag{4.26}$$

Eine einfache lineare Regressionanalyse ergibt für die Gleichung 4.26 einen durchschnittlichen Wert von $f_{exp} = 1.2$. Um einen analytischen Ausdruck dafür zu erhalten, wird ein kleiner Umweg gegangen. Die Exponentialverteilung fällt sehr schnell gegen Null. Für jeden Rauschwert gibt es daher einen Bereich in dem die sich die ursprüngliche Verteilung der Nullfunktion annähert. Nullfunktion heißt: jeder Wert der Urverteilung ist Null. Dort erhält man nach der Addition des Rauschens erwartungsgemäß symmetrische Rang-Intensitäts-Kurven (Abb.4.30.a). Auch erhält man eine Korrelation der eingegangenen Standardabweichung mit dem Mittelwert der negativen Werte (Abb.4.30.b). Aufgrund der Symmetrie des normalverteilten Rauschens ist dieser Mittelwert bei ursprünglicher Nullverteilung gleich dem Mittelwert der positiven Werte. Es läßt sich leicht zeigen, daß diese Mittelwerte in diesem Fall gleich dem Mittleren Absolutfehler η sind. Im exponentiellen Fall ist die diese Symmetrie gestört. Nur der negative Bereich zeigt hier eine Ähnlichkeit mit der Rangverteilung der verrauschten Nullfunktion.

$$\left| \frac{\sum x_-}{n_-} \right| \cong \frac{\sum x_+}{n_+} \quad \cong \quad \eta = \frac{\sum |x|}{n} \tag{4.27}$$

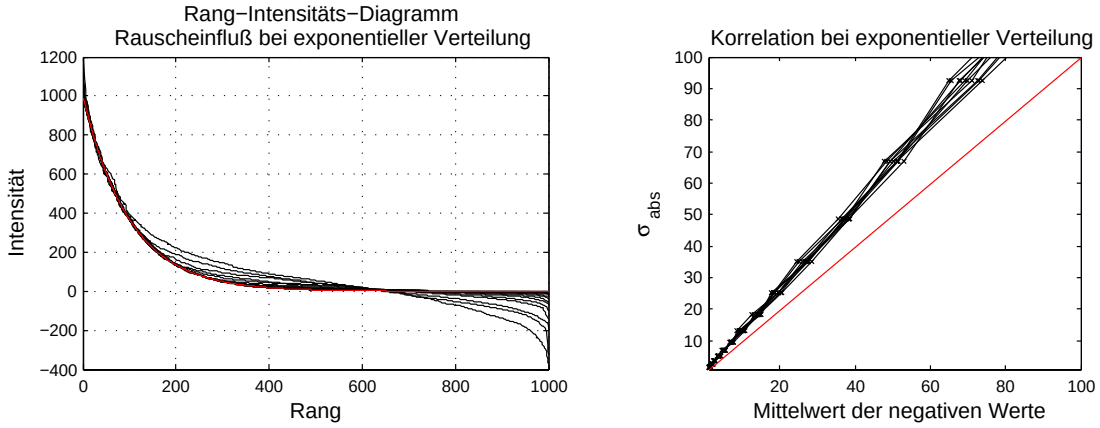


Abbildung 4.27: ***(a)*** *Rang-Intensitäts-Diagramm von Testdaten mit zugrunde liegender* ***Exponentialverteilung*** *(rote Kurve); die schwarzen Kurven repräsentieren diese Verteilung plus einem additiven Rauschen mit definierten Standardabweichung σ_{abs}* ***(b)*** *Korrelation des absoluten Mittelwertes der negativen Werte $\bar{x}_-$ mit der Standardabweichung des addierten Rauschens σ_{abs}. Die rote Gerade entspricht der Identität. Die schwarzen Kurven stellen jeweils ein Set an verschiedenem Rauschen dar. Jede einzelne Kurve ist dabei eine Wiederholung dieser Additionen mit neuer Randomisierung.*

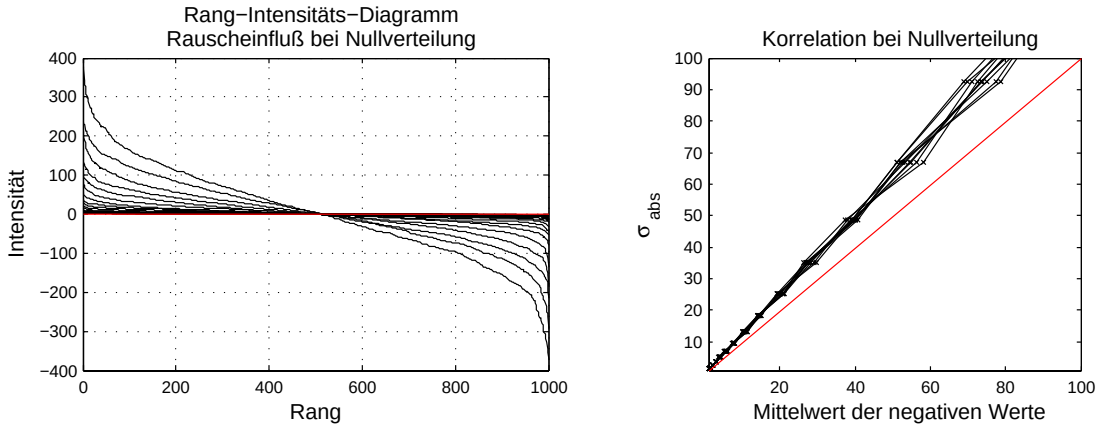


Abbildung 4.28: ***(a)*** *Rang-Intensitäts-Diagramm von Testdaten mit zugrunde liegender* ***Nullverteilung*** *(rote Kurve); die schwarzen Kurven repräsentieren diese Verteilung plus einem additiven Rauschen mit definierten Standardabweichung σ_{abs}* ***(b)*** *Korrelation des absoluten Mittelwertes der negativen Werte $\bar{x}_-$ mit der Standardabweichung des addierten Rauschens σ_{abs}. Die rote Gerade entspricht der Identität. Die schwarzen Kurven stellen jeweils ein Set an verschiedenem Rauschen dar. Jede einzelne Kurve ist dabei eine Wiederholung dieser Additionen mit neuer Randomisierung.*

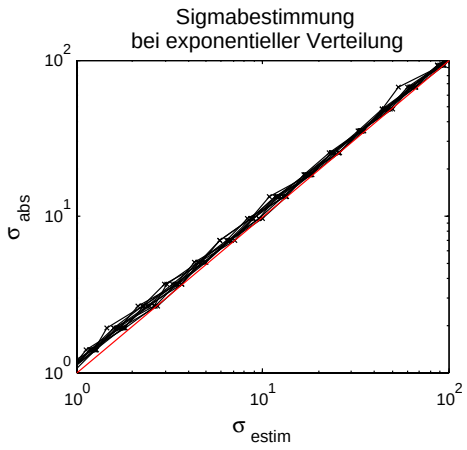


Abbildung 4.29: Logarithmische Darstellung der Korrelation der geschätzten Standardabweichung σ_{estim} mit der wirklichen Standardabweichung des addierten Rauschens σ_{abs} bei zugrunde liegender ***Exponentialverteilung****. Die rote Gerade entspricht der Identität. Die schwarzen Kurven stellen jeweils ein Set an verschiedenem Rauschen dar. Jede einzelne Kurve ist dabei eine Wiederholung dieser Additionen mit neuer Randomisierung.*

Es gibt einen analytischen Zusammenhang zwischen dem Mittleren Absolutfehler η und der Standardabweichung σ ([Bronstein] S.783ff).

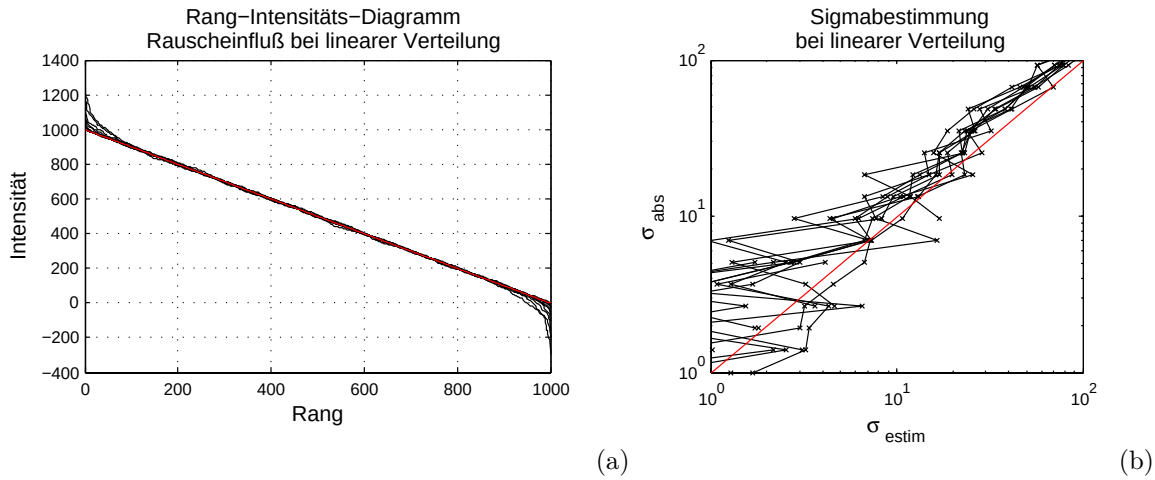
$$\sigma = \sqrt{\tfrac{\pi}{2}}\eta \tag{4.28}$$

$$f_\eta = \sqrt{\tfrac{\pi}{2}} \cong 1.25331\ldots \tag{4.29}$$

Der angegebene Korrelationsfaktor f_η der Gleichung 4.28 ist nahezu identisch mit dem Korrelationsfaktor f_{exp} der exponentiellen Verteilungen. Was sich auf die Ähnlichkeit der negativen Bereiche der Rang-Intensitäts-Kurven zurückführen läßt. Abbildung 4.29 zeigt, daß die resultierende Gleichung 4.30 eine sehr gute Schätzung des ursprünglichen Wertes der Standardabweichung erlaubt.

$$\sigma_{estim} = \sqrt{\tfrac{\pi}{2}}\left|\frac{\sum x_-}{n_{x-}}\right| \tag{4.30}$$

Allerdings ist die Güte der Schätzung abhängig von der ursprünglichen Verteilungsfunktion und der Anzahl der Meßpunkte (Spots). Bei einer ursprünglich linearen Verteilungsfunktion führt Gleichung 4.30 zu einer systematischen Unterschätzung des Fehlers (Standardabweichung). Außerdem führt ein geringes addiertes Rauschen nur zu wenigen negativen Werten. Je geringer die Anzahl der negativen Meßpunkte ist, desto unsicherer ist die Schätzung (Abb. 4.30.b). Auch bei der Fehlerschätzung der Exponentialverteilung erfolgt ein leichte Unterschätzung. Mittels linearer Regression ließe sich noch eine bessere Anpassung vornehmen. Doch in den Realdaten ist die wahre Urverteilung unbekannt, so daß diese selbst geschätzt werden muß. Bei einer relativen Unterschätzung von ≈5% der wirklichen Standardabweichung mit obiger Gleichung ist das nicht unbedingt notwendig. Wenn doch, kommt als Näherung der Urverteilung z.B. die Verteilung des Experimentes mit dem geringsten Fehler in Frage.



(a) (b)

Abbildung 4.30: ***(a)*** *Rang-Intensitäts-Diagramm von Testdaten mit zugrunde liegender* ***linearer Verteilung*** *(rote Kurve); die schwarzen Kurven repräsentieren diese Verteilung plus einem additiven Rauschen mit definierten Standardabweichung σ_{abs}* ***(b)*** *logarithmische Darstellung der Korrelation der geschätzten Standardabweichung σ_{estim} mit der wirklichen Standardabweichung des addierten Rauschens σ_{abs}. Die rote Gerade entspricht der Identität. Die schwarzen Kurven stellen jeweils ein Set an verschiedenem Rauschen dar. Jede einzelne Kurve ist dabei eine Wiederholung dieser Additionen mit neuer Randomisierung.*

4.4 Vergleich der Normalisierungsmethoden und ihrer Fehler

4.4.1 Motivation

Wie in Kapitel 2 beschrieben müssen alle Normalisierungsmethoden verschiedene Grundannahmen über diverse Dateneigenschaften machen. Die Gültigkeit der Annahmen ist im Wesentlichen von der Fragestellung, der Meßmethode und den letztendlichen Daten abhängig. Doch auch unabhängig davon, kann die jeweilige Methode zusätzliche Fehler einführen oder verstärken. Dieser Fehlereinfluß stellt, neben der Optimierung an die jeweilige Grundannahme, ein zusätzliches Kriterium für die Auswahl der günstigsten Normalisierung dar. Im folgenden werden verschiedene Normalisierungsmethoden mittels dieser Kriterien verglichen.

4.4.2 Fehlerkriterium

Nimmt man an, daß die jeweilige Grundannahme gerechtfertigt ist, dann ist die Normalisierungsmethode die beste, die den geringsten Fehler in die Daten bringt. Der Fehlereinfluß durch die Transformationsparameter der Normalisierung ergibt sich durch die Fehlerfortpflanzung aus der Transformationsgleichung. Allgemeine Transformationsgleichung:

$$s_n = \frac{s_p - \gamma}{\kappa} \tag{4.31}$$

Maximalfehler des transformierten (normalisierten) Wertes:

$$\overline{\Delta s_n} = \left|\frac{1}{\kappa}\right| \overline{\Delta s_p} + \left|\frac{\gamma - s_p}{\kappa^2}\right| \overline{\Delta\kappa} + \left|-\frac{1}{\kappa}\right| \overline{\Delta\gamma} \tag{4.32}$$

relativer Maximalfehler des transformierten (normalisierten) Wertes:

$$\frac{\overline{\Delta s_n}}{s_n} = \frac{\overline{\Delta s_p} + \overline{\Delta\gamma}}{|s_p - \gamma|} + \frac{\overline{\Delta\kappa}}{|\kappa|} \tag{4.33}$$

Die Gleichungen gelten allgemein. Für globale lineare Methoden ist dieses offensichtlich. Die beiden Transformationsparameter κ, der Skalierungsparameter, und γ, der Biasparameter, und ihre Fehler sind dort für alle Werte einer Messung konstant. Nichtlineare Methoden kann man verallgemeinert definieren, als Transformationen bei denen die Parameter nicht global für ein Experiment gelten, sondern lokal abhängig vom zu normalisierenden Wert sind.

Die wichtigere Gleichung ist die des relativen Fehlers, da die Daten durch die Normalisierung relativ zu einem Standard normiert werden. Aus der Gleichung folgt, daß für die Minimierung dieses relativen Fehlers: der Absolutfehler der Biasparameters $\Delta\gamma$ und der Relativfehler des Skalierungsparameters $\frac{\overline{\Delta\kappa}}{\kappa}$ minimiert werden müssen.

Die meisten Methoden gehen von einer korrekten Hintergrundbestimmung aus. Der Biasfehler ist hier schon im Fehler des Rohwertes enthalten. Damit vereinfacht sich die Gleichungen 4.31 und 4.33 zu:

$$s_n = \frac{s_p}{\kappa} \tag{4.34}$$

$$\frac{\overline{\Delta s_n}}{s_n} = \frac{\overline{\Delta s_p}}{|s_p|} + \frac{\overline{\Delta\kappa}}{|\kappa|} \tag{4.35}$$

4.4.3 Verteilungskriterium

Im Abschnitt 4.2 wurde die relative Rang-Intensitäts-Standardabweichung (rRISA) als Beschreibungsgröße für die globale Abweichung der rangbasierte Intensitäts-Verteilung (RIV) innerhalb eines Experimentvergleiches eingeführt. Als Bezugsverteilung dient hier die Mittelwertkurve aller zu vergleichenden Experimente. Es ist nun möglich rRISA als Maß der Vergleichbarkeit der Verteilungen zu definieren. Verschiedene systematische Fehler beeinflußen, wie oben gezeigt, die RIV. Je ähnlicher die RIV sind, desto besser sollten die Daten vergleichbar sein. (Ähnlich heißt hier, die Kurven können im wesentlichen durch einfache Transformation in einander überführt werden.) Es liegt daher nahe Funktionen, die die Verteilung beeinflußen (wie die Normalisierung) auch an der Minimierung der Verteilungsabweichung (rRISA) zu messen.

4.4.4 Testdaten

Wie schon im Abschnitt zuvor werden zum Testen der Normalisierungsmethoden Testdaten eingesetzt. Als Urverteilung kommt wiederum eine exponentielle Verteilung zum Einsatz (Siehe Gl.3.2 auf S. 23). Damit ist in den Testdaten die rauschfreie Urverteilung bekannt. Es wird angenommen, daß kein Biasfehler auftritt und daß nur ein additives normalverteiltes Rauschen auftritt. Dieses wird jeweils neu randomisiert und auf die Urverteilung addiert.

4.4.5 lineare Referenzmethoden

Interne Referenzgene

Die einfachste Normalisierung ist die Verwendung eines einzelnen Genes als Skalierungsparameter (Gl.4.36). Unter der Annahme der Konstanz dieses Parameters in jedem einzelnen Experiment, bringt dieser den Fehler der Einzelmessung ein (Gl.4.37).

$$\kappa = s_{p,ref} \tag{4.36}$$

$$\frac{\overline{\Delta\kappa}}{\kappa} = \frac{\left|\overline{\Delta s_{p,ref}}\right|}{s_{p,ref}} \tag{4.37}$$

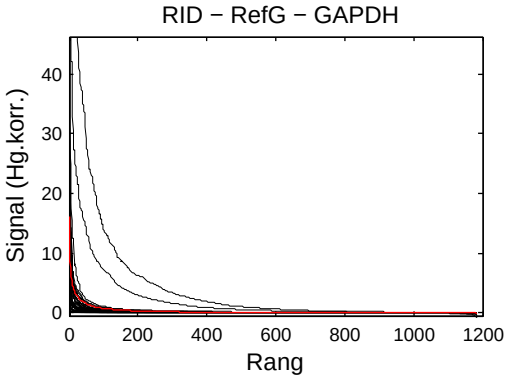


Abbildung 4.31: ***GAPDH*** *normalisierte Rang-Intensitäts-Kurven von HAA1.2 Experimenten*

Daraus folgt, bei gleichem Absolutfehler der Messung über den gesamten Meßbereich, daß ein stark exprimiertes Referenzgen besser zur Normalisierung geeignet ist als ein niedrigeres. (Allerdings können Sättigungseffekte dieser Tendenz entgegen wirken, da sie einen höheren Meßwertfehler einführen)

Internes Referenzgenset

Günstiger in dieser Beziehung ist die Verwendung mehrerer Referenzgene (Gl.4.38), da durch die Mittelwertbildung der Fehlereinfluß der Messung verringert wird. Wenn für alle Referenzen der gleiche Fehler gilt, kann man den resultierenden Fehler über den Fehler des arithmetischen Mittelwertes bestimmen (Gl.4.39). Für den Relativfehler gilt dann Gleichung (Gl.4.40).

$$\kappa = \bar{s}_{p,ref} = \frac{\sum_{i=1}^{N_{ref}} s_{p,ref,i}}{N_{ref}} \tag{4.38}$$

$$\overline{\Delta \bar{s}_{p,ref}} = \frac{\overline{\Delta s_p}}{\sqrt{N_{ref}}} \tag{4.39}$$

$$\frac{\overline{\Delta \kappa}}{\kappa} = \frac{\overline{\Delta \bar{s}_{p,ref}}}{\bar{s}_{p,ref}} = \frac{\overline{\Delta s_p}}{\bar{s}_{p,ref} \cdot \sqrt{N_{ref}}} \tag{4.40}$$

Wenn die Fehler unterschiedlich sind erfolgt die Berechnung über folgende Gleichung 4.41.

$$\frac{\overline{\Delta \kappa}}{\kappa} = \frac{\left|\overline{\Delta \bar{s}_{p,ref}}\right|}{\bar{s}_{p,ref}} = \frac{\sqrt{\sum_{i=1}^{N_{ref}} \Delta s^2_{p,ref,i}}}{N_{ref} \cdot \bar{s}_{p,ref}} \tag{4.41}$$

Realdaten Bei der Anwendung der Normalisierung durch Referenzgene auf Realdaten, z.B. Daten von HAA1.2, kann eigentlich die Verwendung des Verteilungskriteriums nur bedingt herangezogen werden, da die Grundannahme der Konstanz der Referenzgene nicht äquivalent der Grundannahme des Verteilungskriteriums ist. Somit können die nachfolgenden rangbasierten Diagramme nur die globalen Auswirkungen dieser Normalisierungsmethode demonstrieren. Sie belegen nicht die Güte der jeweiligen Grundannahme. Als Beispiel dient das Standardreferenzprotein GAPDH. In Diagramm 4.31 wird das Ergebnis dieser Normalisierung an Hand des RID verschiedener Experimente gezeigt.

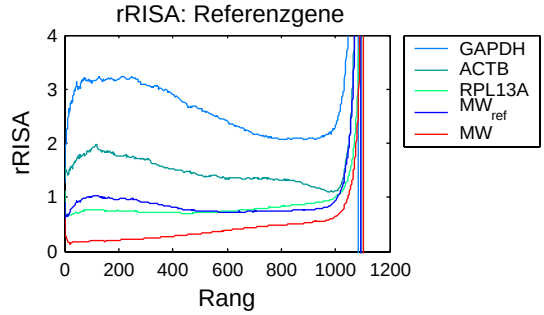


Abbildung 4.32: rRISA-Diagramm: Vergleich der rRISA der Normalisierung durch interne Referenzgene mit der durch den globalen Mittelwert(MW) Realdaten von HAA1.2

Der Vergleich der Referenzmethoden mittels der rRISA ergibt folgendes Bild (Abbildung 4.32): Bezüglich des Kriteriums ergibt keines der Referenzgene eine bessere Angleichung der Kurven als die Mittelwertnormalisierung. Die Kurven der beispielhaft ausgewählten Gene GAPDH[1] und ACTB[2] haben eher eine Ähnlichkeit mit dem Resultat der Max-Normalisierung (siehe Perzentile), was darauf schließen läßt, daß ihre Signale durch Sättigungseffekte beeinflußt sein könnten (zusätzlicher systematischer Fehler). Dadurch, daß die meisten Referenzen zu den stärksten Signale auf dem Array gehören, zeigt der Mittelwert aus den Referenzgenen das gleiche Verhalten.

4.4.6 lineare Globalisierungsmethoden

Alle Globalisierunsmethoden basieren auf der Annahme, (i) daß die Menge an mRNA pro Zelle konstant ist und (ii) daß die Menge an hybridisierter mRNA proportional der Gesamtmenge an mRNA ist.

Mittelwert

Bei der Mittelwertnormalisierung wird der globale Mittelwert aller Signale als Normalisierungsparameter verwendet (Gl. 4.42). Wie auch die folgenden ist diese Methode eine einfache Skalierung. Der Fehler durch die Normalisierung ergibt sich aus den Gleichung 4.43. Ist der Fehler der Einzelwerte gleich, ist der Fehler des arithmetischen Mittelwertes durch Gl. 4.44 gegeben. Die Skalierung mit dem Mittelwert ist die Standardmethode bei der Globalisierung. Die Bestimmung des Mittelwertes wird durch nichtlineare Effekte in den extremen Wertbereichen (höchste und niedrigste) beeinflußt.

$$\kappa = \bar{s}_p = \frac{\sum_{i=1}^{N} s_{p,i}}{N} \tag{4.42}$$

$$\frac{\overline{\Delta\kappa}}{\kappa} = \frac{\left|\overline{\Delta\bar{s}_p}\right|}{\bar{s}_p} = \frac{\sqrt{\sum_{i=1}^{N} \Delta s_{p,i}^2}}{N \cdot \bar{s}_p} \tag{4.43}$$

$$\overline{\Delta\bar{s}_p} = \frac{\overline{\Delta s_p}}{\sqrt{N}} \tag{4.44}$$

[1] Glyzeraldehyd-3-Phosphatdehydrogenase
[2] β -Actin

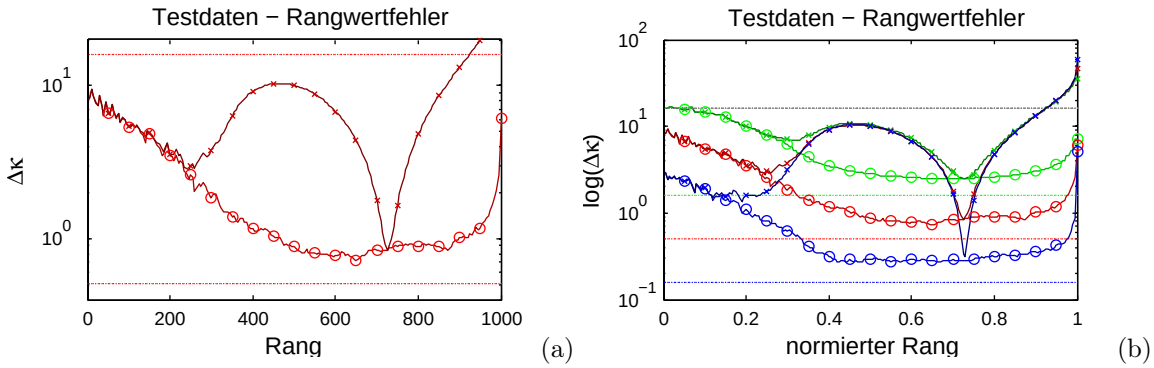


Abbildung 4.33: ***(a)Rangwertfehler der Testdaten:*** *Die rote −o− Kurve stellt die Standardabweichung zwischen 100 Rang-Intensitäts-Kurven mit gleichem Rauschen dar ($\sigma = 16$). Es ist ein Maß für den Fehler des Rangwertes. Die rote −x− Kurve stellt die Standardabweichung dieser Kurven von der Urverteilung dar. Die gestrichelten Linien sind Vergleichsmaße. Die obere ist die Standardabweichung des Einzelwertes σ . Die untere ist der Fehler des Mittelwertes (N=1000).* ***(b)Abhängigkeit des Rangwertfehler von der Anzahl der Meßpunkte N:*** *Allen betrachteten Kurven liegt die selbe Urverteilung und dasselbe Rauschen zugrunde. Jede gleichfarbige Gruppe entspricht dem Diagramm (a) (grün: N=100, rot: N=1000, blau N=10000. Allerdings wurde der normierte Rang als Abszisse verwendet, damit sich die Rang-Intensitäts-Kurven bei unterschiedlicher Rangzahl miteinander vergleichen lassen. Die graue Linie stellt die den Kurven gemeinsame Standardabweichung dar σ*

Realdaten Die Anwendung der mittelwertbasierten Normalisierung auf die Beispieldaten (HAA1.2) wurde schon bei der Einführung des Rang-Intensitäts-Diagrammes gezeigt und beschrieben (Abb. 4.26 auf S. 52).

Quantile/Perzentile

Die folgende Methode basiert auf der Ähnlichkeit der Rang-Intensitäts-Verteilung. Es wird vermutet, daß Einflüsse auf die Extremwertbereiche, keinen Einfluß auf die zentrale Werteverteilung haben, so daß bei sonstiger Ähnlichkeit eine relative Skalierung auf einen der Rangwerte[3] die Kurven aneinander anpassen sollte. Normiert man die, dem Wert zugeordneten Ränge auf die Gesamtanzahl, werden die Werte Quantile genannt (oder Perzentile bei der Normierung auf 100%). Quantile die nicht direkt einem Rangwert zugeordnet werden, können über lineare Interpolation zwischen den beiden nächsten Rangwerten ermittelt werden.

Die Vermutung der weitgehenden Verteilungsgleichheit gilt, wenn überhaupt, nur für genomische oder repräsentative Arrays. Obwohl einzelne extremregulierte Gene bei geeigneter Parametrisierung einen geringeren Einfluß haben, als bei der Mittelwertnormalisierung.

Die Rang-Intensitäts-Verteilung wird weiterhin durch Rauscheinfluß verändert. Im vorangehenden Abschnitt wurde die Möglichkeit der Rauschabschätzung aufgrund dieses Verhaltens beschrieben. Diese Veränderung ist in den unteren Wertebereichen proportional zum Ausmaß des Rauschens. Wird nun in diesem Bereich auf einen enthaltenen Rangwert/Quantil skaliert, repräsentiert der Skalierungswert κ nicht mehr die Urverteilung, sondern vorwiegend den Rauscheinfluß (Abb. 4.27). Daher muß bei Verwendung dieser Methode, der letzendliche benutzte Quantil aus einem möglichst gering durch Rauschen beeinflußten Bereich stammen. Um diesen Bereich zu bestimmen, kann das RISA-Diagramm benutzt werden. In diesem werden die absoluten Abweichungen von der Urverteilung dargestellt.

[3]Der Rangwert ist der größengeordnete Wert, der einem bestimmten Rang (Ordnungszahl) entspricht.

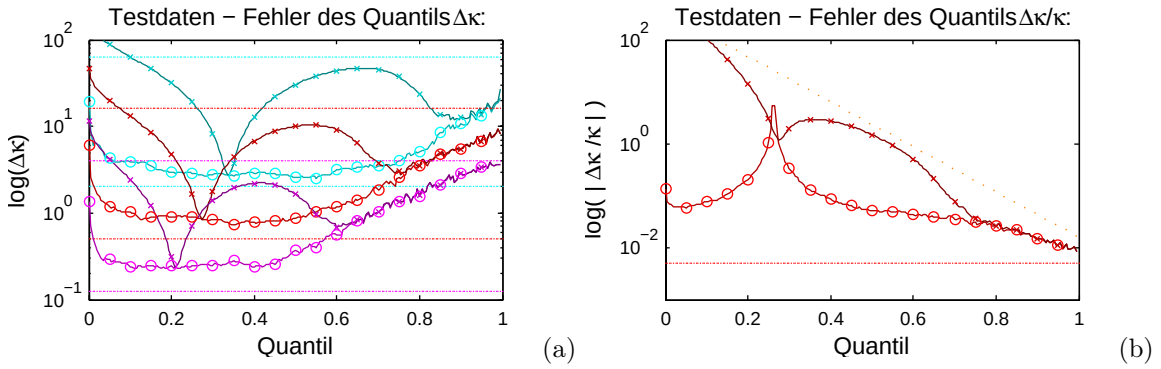


Abbildung 4.34: ***(a) absoluter Fehler des Quantils:*** *Das Diagramm ist ähnlich zu Abb. 4.33. Im Unterschied dazu ist auf Abszisse der Quantilparameter aufgetragen. Die Ordinate ist logarithmisch um die Auswirkungen dreier unterschiedlicher Standardabweichungen auf die Varianz der Rangwertverteilung darzustellen. Die $-o-$ Kurven stellen die Standardabweichung zwischen 100 Rang-Intensitäts-Kurven mit gleichem Rauschen dar ($\sigma_{trkis} = 32$,$\sigma_{rot} = 16$,$\sigma_{magenta} = 8$). Die $-x-$ Kurven stellen die Standardabweichung dieser Kurven von der Urverteilung dar. Die gestrichelten Linien sind Vergleichsmaße. Die obere ist die Standardabweichung des Einzelwertes σ . Die untere ist der Fehler des Mittelwertes (N=1000).* ***(b) relativer Fehler des Quantils:*** *Im Gegensatz zu den vorherigen Abbildungen ist hier der relative Fehler des Quantilwertes aufgetragen. Die logarithmische Auftragung erlaubt dabei eine bessere Darstellung über den gesamten Wertebereich. Die $-o-$ Kurve zeigt die approximierte relative Standardabweichung des Quantils. Die $-x-$ Kurve ist die relative Abweichung des Quantilwerts von der Urverteilung. Die gestrichelten Linien sind Vergleichsmaße. Die obere orangene ist die wahre relative Standardabweichung des Einzelwertes $\frac{\overline{\Delta s_p}}{|s_p|}$ (in Reihenfolge des zugehörigen Quantilwerts). Die untere rote ist der Fehler des Mittelwertes (N=1000).*

Wie kann nun der Fehlereinfluß des jeweiligen Quantils abgeschätzt werden? Dazu werden wieder die Testdaten mit der exponentielle Verteilung herangezogen. Um den Fehler der Rangwerte zu schätzen, wird die Standardabweichung zwischen mehreren rauschbeeinflußten Verteilungen bestimmt. Dabei wurde zu jeder Kurve ein Rauschen gleicher Standardabweichung addiert. Das RISA-Diagramm (Abb. 4.33.a) stellt die Standardabweichung zwischen den Kurven dar. Die obere rote Linie stellt die ursprüngliche Standardabweichung des addierten Rauschens dar. Wie man sehen kann ist die Standardabweichung zwischen den Kurven viel geringer als diese. Daraus folgt, daß der Fehler des Rangwertes (Quantils) geringer ist als der Fehler des Einzelwertes. Dieser Effekt wird durch die Ordnungsprozedur der Verteilungskurve erreicht. Je mehr Werte in der Verteilungskurve enthalten sind ($\widehat{=}$ mehr Sonden auf einem Array), desto geringer ist der Fehler des Rangwertes (Abb. 4.33.b). Für Quantile die keine direkte Rangwertzuordnung haben, kann der Fehler durch lineare Interpolation aus den Fehlern der beiden nächsten Rangwerten geschätzt werden.

Es zeigt sich also, daß das Rauschen zu systematischen Veränderungen der Rang-Intensitäts-Verteilung führt, die allerdings sehr konsistent sind (Abb. 4.34.a). In Realdaten ist die Urverteilung nicht bekannt, und es sind weitere Einflußgrößen vorhanden, somit läßt sich dieser systematische Einfluß nicht herausrechnen. Man hat also zwei gegenläufige Effekte: Je geringer der Rangwert ist, desto weniger schwankt der Wert und je größer der Rangwert, desto geringer ist der systematische Einfluß des Rauschen. Allerdings geht in die Normalisierung der relative Fehler ein und so vereinfacht sich für die Testdaten die Aussage zu: Je größer der Rangwert (Quantil), desto geringer ist der Normalisierungsfehler durch dessen Verwendung als Skalierungsquotient (Abb. 4.34.b).

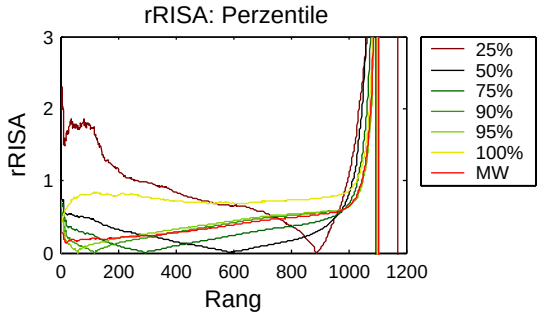


Abbildung 4.35: ***rRISA-Diagramm:*** *Vergleich der relativen Standardabweichung der Rang-Intensitäts-Kurven (rRISA) der verschieden Perzentil - Normalisierung mit dem Mittelwert (MW) Realdaten von HAA1.2*

Realdaten Die Abbildung 4.35 zeigt die rRISA-Kurven der Normalisierung mit höheren Perzentilen. Als empirische Näherung für die unbekannte Urverteilung wird das Rangwertmittel verwendet. Wie leicht zu sehen ist, kann man durch die Verschiebung des Perzentil eine unterschiedliche Angleichung der Kurven aneinander erreichen. Höhere Perzentile führen bis etwa 95% führen zu einer Verbesserung der Anpassung. Ein höherer Wert (100%-Max) führt wiederum zu einer stärkeren rRISA der Kurven. Diese ist wahrscheinlich durch stärkere Streuung der Verteilung im Bereich der sehr hohen Werte verursacht (z.B. Sättigungseffekte oder Überexpression einiger weniger Gene in einem kleinen Genset).

Asymmetrisch gestutztes Mittel

Wie oben für die Referenzgene gezeigt, führt die Verwendung des Mittelwertes mehrerer Referenzen zu einer Reduzierung des Fehler des Skalierungsquotienten. Ähnlich dazu könnte man mehrere Quantile zusammenfassen oder den Mittelwert eines Rangbereiches benutzen. Die Verwendung eines Mittelwertbereiches führt zu der Methode des gestutzten Mittels, da der Bereich der Mittelwertbildung um einen unteren und oberen Bereich gestutzt wird. Die bisher übliche Methode des gestutzten Mittels geht dabei von einer symmetrischen Stutzung aus [Tukey1962]. Das Ziel dieser Methode ist, einen Ausreißer resistenten Schätzer für den wahren Wert einer Mehrfachmessung unter dem Einfluß normalverteilten Rauschens zu haben. Das würde etwa der Nullverteilung im vorherigen Abschnitt entsprechen (Siehe S.56). Bei dieser liegt eine weitgehend symmetrische Kurve vor. In Falle der Genexpressionsdaten ist diese Verteilung aber nicht symmetrisch, sondern mit einer exponentiellen Verteilung überlagert. Die Bereiche der hohen (oberen) und niedrigen (unteren) Extremwerte sind daher, wie mehrfach beschrieben, unterschiedlich vom Rauschen beeinflußt. Weiterhin hat eine obere Stutzung einen stärkeren Einfluß auf den Mittelwert als eine im unteren Bereich. Eine unbedingte symmetrische Stutzung ist daher in diesem Fall nicht zu begründen.

Es wird hier das asymmetrisch gestutzte Mittel (AGM) eingeführt, als Methode um einen Mittelwert eines beliebigen Bereiches einer Rang-Intensitäts-Verteilung zu bestimmen [Kroll2002B]. In diesem Fall ist dieser Mittelwert ein Schätzer des wahren relativen Skalierungsquotienten, unterschiedlich skalierter und verrauschter Rang-Intensitäts-Verteilungen mit gleicher Urverteilung (Normalisierungsannahme !). Das AGM besitzt zwei Parameter: den oberen Stutzungsanteil o(in Teilen von 1 oder 100%) und den unteren Stutzungsanteil u. Die Summe aus beiden Stutzungsparametern muß kleiner 1 oder 100% sein. Die Stutzungsparameter entsprechen Quantilen. Für die Mittelwertbildung wird die rangnormierte Rang-Intensitäts-Kurve zwischen dem $(1-o)$-ten Quantil und dem u-ten Quantil integriert. Wobei die Kurve zwischen den ganzrängigen Quantilen[4] linear interpoliert ist. Diese Summe wird durch den $(1-o-u)$-ten Rangbereich dividiert (Gl.4.45). Die Umsetzung dieses Algorithmus erfolgte in MATLAB (siehe Anhang S.VII).

[4]Quantil der direkt ohne Rundung einem Rang zugeordnet werden kann

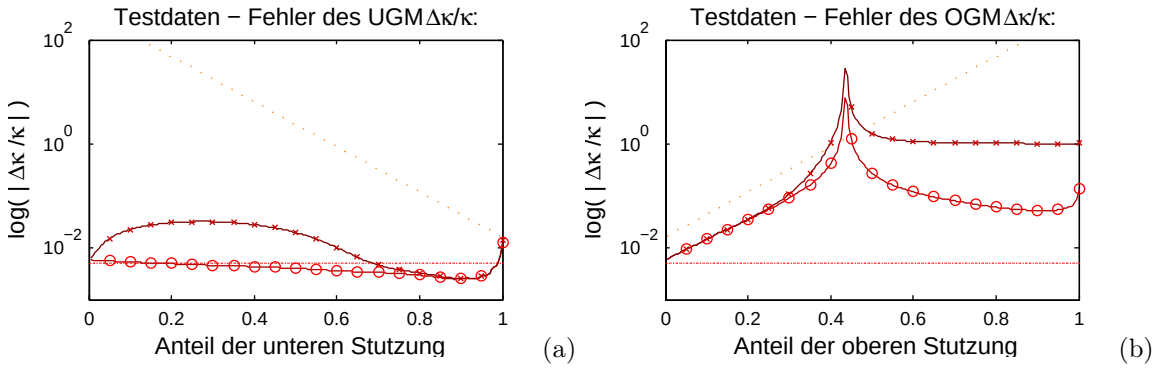


Abbildung 4.36: ***relativer Normalisierungsfehler*** *Die beiden Diagramme sind äquivalent zu Abb. 4.34.b . Die Abszisse stellt jeweils den Stutzungsparameter des* ***(a) unteren und (b) oberen gestutzten Mittels*** *dar. Die $-o-$ Kurve zeigt die approximierte relative Standardabweichung der Methode. Die $-x-$ Kurve ist die relative Abweichung des gestutzten Mittelwerts von der Urverteilung. Die gestrichelten Linien sind Vergleichsmaße. Die obere orangene ist die wahre relative Standardabweichung des Einzelwertes $\frac{\overline{\Delta s_p}}{|s_p|}$ (in (a) umgekehrter Reihenfolge und (b) in Reihenfolge des Normranges). Die untere rote ist der Fehler des Mittelwertes (N=1000).*

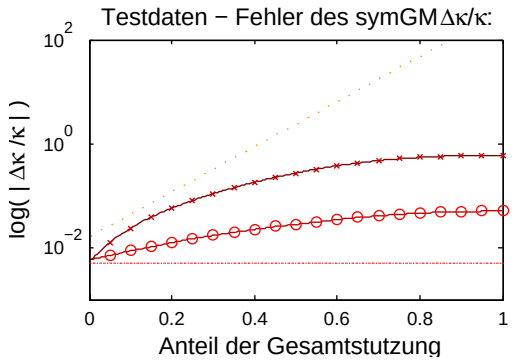
$$\kappa \quad = {}^{o}_{u}\bar{s} = \quad \frac{\int_u^o \text{quantil}_q(s_{p,all})dq}{(1-o-u)N} \tag{4.45}$$

Eine vereinfachte Form ist die diskrete Umsetzung des AGM mit der Rundungsfunktion ohne Interpolation zwischen den Rangwerten in Gleichung 4.46).

$$\kappa = {}^{o}_{u}\bar{s} \quad = \quad \frac{1}{\text{int}(\frac{1}{2} + (1-o-u)N)} \sum_{r=\text{int}(\frac{1}{2}+N*o)}^{\text{int}(\frac{1}{2}+N(1-u))} s_r \tag{4.46}$$

Um den Einfluß der beiden Stutzungsparameter u und o auf den Fehler des AGM zu bestimmen, werden wie oben mehrere Testexperimente mit gleichem Rauschen verwendet und die berechnete relative Standardabweichung zwischen den rauschbeeinflußten AGM gleicher Stutzung als Schätzer der Standardabweichung benutzt. Abbildung 4.36.a stellt den relativen Fehler $\frac{\overline{\Delta\kappa}}{\kappa}$ des unteren gestutzten Mittels (UGM) dar. Wobei auf der Abszisse der untere Stutzungsparameter u aufgetragen ist. Die Abbildung zeigt, daß das untere Stutzen bezüglich Standardabweichung des gestutzten Mittels zu einer Verringerung desselben führt (rote $-o-$ Kurve). Allerdings zeigt sich ein starke Abweichung von den zugehörigen gestutzten Mittel der Urverteilung ($-x-$ Kurve). Daraus folgt, daß durch das Stutzen im unteren Bereich der systematische Einfluß des Rauschens auf den Normalisierungsquotienten verstärkt wird. Erst bei einer sehr starken Stutzung von etwa 0.8 (80%) hat sich dieser Einfluß niveliert. Eine noch stärkere Stutzung führt sogar zu einer Reduktion beider Fehler gegen über dem normalen relativen Mittelwertsfehler. Wenn die Stutzung gegen 1 (100%) geht, verstärkt sich der Fehler und er nähert sich dem Fehler des Maximalwertes (100% Perzentil), d.h. für die Testdaten gibt es ein Optimum der unteren Stutzung von etwa 0.9 (90%).

Der andere Spezialfall des AGM, das obere gestutzten Mittel, (OGM) ist in Abbildung 4.36.b dargestellt. Hier führt jede Stutzung zu einer Verschlechterung des Fehlers. Die Polstelle bei einer oberen Stutzung von 0.4 (40%) ist darauf zurückzuführen, daß der Mittelwert dieses Rangwertbereiches negativ wird. Interessanterweise weicht die Kurve der relativen Abweichung von der Urverteilung ($-x-$) bis zu diesem Punkt nicht von der relativen Standardabweichung des zugehörigen OGM-Wertes ab($-o-$).



(a)

Abbildung 4.37: ***relativer Normalisierungsfehler*** *Diagramm ist äquivalent zu Abb. 4.34.b . Die Abszisse stellt den Gesamtstutzungsparameter des* ***symmetrischen Mittels*** *dar. Die −o− Kurve zeigt die approximierte relative Standardabweichung der Methode. Die −x− Kurve ist die relative Abweichung des gestutzten Mittelwerts von der Urverteilung. Die gestrichelten Linien sind Vergleichsmaße. Die obere orangene ist die wahre relative Standardabweichung des Einzelwertes* $\frac{\overline{\Delta s_p}}{|s_p|}$ *(in (a) umgekehrter Reihenfolge und (b) in Reihenfolge des Normranges). Die untere rote ist der Fehler des Mittelwertes (N=1000).*

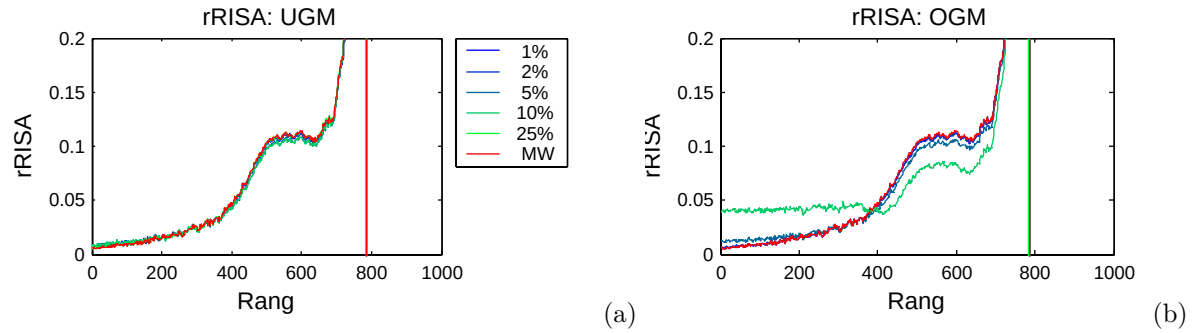
Es zeigt sich das die Stutzungen unabhängig von einander sind. Untere Stutzung erscheint sinnvoll. Obere Stutzung verschlechtert den Fehler des gestutzten Mittels. das heißt aber auch, daß die symmetrische Stutzung der klassische Methode des gestutzten Mittels (symGM) nicht für diese Art der Verteilung geeignet ist. Allerdings ist der Unterschied zum reinen oberen Stutzen nicht sehr groß (Abb.4.4.6.a). Beim Vergleich der Abbildungen der beiden Methoden muß beachtet werden, daß ein Stutzen um 0.1 (10%) beim symGM einem oberen Stutzen um 0.05 (5%) entspricht. Beachtet man dieses, ist das symGM bezüglich des relativ Fehlers leicht besser, bezüglich der relativ Abweichung von der Urverteilung aber schlechter.

Es zeigt sich also, daß der Bereich der Mittelwertbildung entweder den gesamten rauschbeeinflußten unteren Bereich einschließen sollte oder ihn gesamtheitlich ausschließen. Allerdings hat die unteren Stutzung auf die eigentliche Normalisierung kaum Einfluß, wie das rRISA-Diagramm eines Testdatensatzes zeigt. (Die verwendeten Testdaten haben die gleiche Urverteilung wie die oben verwendeten. Nur wurden hier 50 Datensätze mit unterschiedlichem Rauschen verwendet um die Realdaten besser zu repräsentieren.)

Die obere Stutzung hat dagegen einen deutlicheren Effekt auch im rRISA-Diagramm. Die Stutzung um 25% zeigt bereits deutlich eine Verschlechterung der Normalisierung. Das ist der Bereich, der Verteilung der nur noch die mittleren und kleinen Werte umfaßt. Das ist auch der Bereich, in dem die Werteverteilung durch das Rauschen verändert wird. Eine zu starke Stutzung führt dann zu einem ähnlichen Effekt wie die Normalisierung mit den unteren Quantilen: der Rauscheinfluß wird verstärkt.

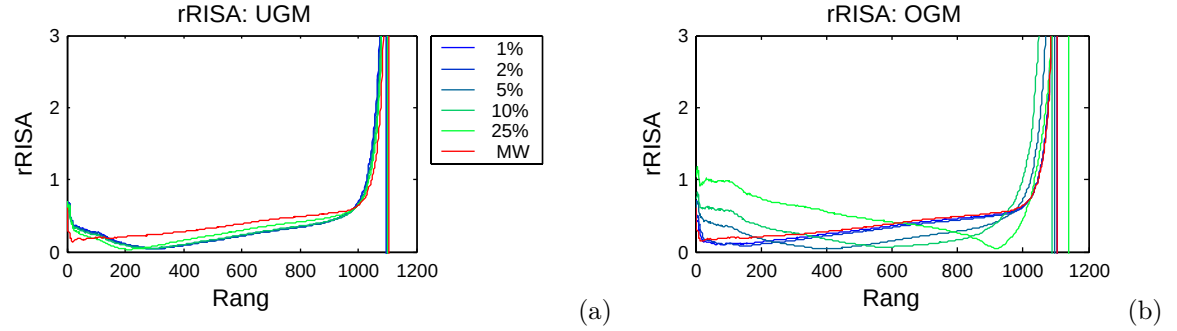
Der Unterschied zwischen oberer und unterer Stutzung liegt im Einfluß der Größe des Wertes des gestutzten Mittels begründet. Durch die untere Stutzung wird dieser Wert nur minimal verändert. Die obere Stutzung beschneidet den hohen Wertebereich: der resultierende Wert wird sehr stark reduziert. Der resultierende relative Fehler $\frac{\overline{\Delta\kappa}}{\kappa}$ ist schon aufgrund der Veränderung des Normalisierungsquotienten κ unterschiedlich von der Stutzung betroffen.

Realdaten Im Gegensatz zu den perfekten Testdaten treffen einige der gemachten Annahmen nicht vollständig auf die Realdaten von HAA1.2-Arrays zu. Wie mehrfach erwähnt, weicht die Werteverteilung besonders bei den sehr hohen Werten zwischen den Experimenten ab. Dieser hohe Bereich besitzt



(a) (b)

Abbildung 4.38: ***rRISA-Diagramm:*** *Vergleich der relativen Standardabweichung der Rang-Intensitäts-Kurven (rRISA) verschiedener Stutzungen des Mittelwerts* ***(a) untere Stutzung (UGM) (b) obere Stutzung (OGM)****. Die Legende gilt für beide Diagramme und gibt die jeweiligen Anteile der Stutzung wieder. Testdaten*



(a) (b)

Abbildung 4.39: ***rRISA-Diagramm:*** *Vergleich der relativen Standardabweichung der Rang-Intensitäts-Kurven (rRISA) verschiedener Stutzungen des Mittelwerts* ***(a) untere Stutzung (UGM) (b) obere Stutzung (OGM)****. Die Legende gilt für beide Diagramme und gibt die jeweiligen Anteile der Stutzung wieder. Realdaten von HAA1.2*

demzufolge einen höheren Rangwertfehler als die restlichen Werte. Wie man an der rRISA-Kurve des Mittelwerts sieht: die obersten 2% der Werte weichen stärker voneinander ab, als die direkt nachfolgenden. Eine Stutzung des Mittelwertbereiches um diese oberen 2% verbessert die Normalisierung nach dem rRISA-Kriterium (OGM Abb. 4.39.a). Eine Stutzung im unteren Bereich hat dagegen kaum Auswirkungen (OGM Abb. 4.39.b). Der Effekt der Veränderung der Verteilung durch das Rauschen, kann aufgrund der unbekannten Urverteilung nicht bewertet werden. Da sich die hier betrachteten Realdaten im unteren Wertebereich ähnlich verhalten wie die Testdaten, kann man auch hier davon ausgehen, daß das untere Stutzen keine Verbesserung der Normalisierung bringt.

4.4.7 nichtlineare Methoden

Die Anpassungen der linearen Normalisierungen versuchen Probleme bei der Bestimmung der Normalisierungsparameter zu umgehen. Dabei werden möglichst Bereiche vermieden, die eine vermutlich nichtlineare Signalfunktion haben. Diese Bereiche können nur mit mäßigen Erfolg normalisiert werden. Abhilfe können nichtlineare Methoden zu liefern, die auch diese Bereiche vergleichbar machen. Die folgende Methode des Rangwertmittels basiert auch auf den Intensitäts-Rang-Verteilungen.

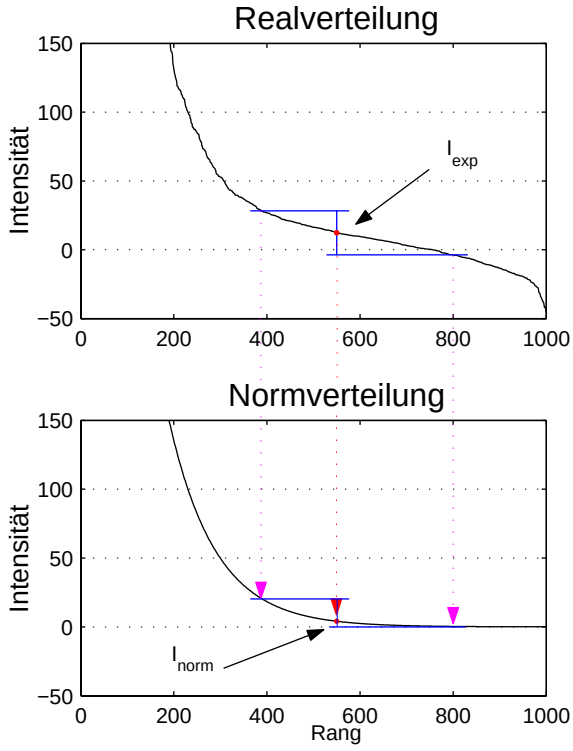


Abbildung 4.40: ***Fehlerfortpflanzung bei Normalisierung mit Referenzverteilung*** *Die obere Verteilung stellt verrauschte Testdaten dar. Die untere Verteilung die zugrunde liegende Urverteilung. Über den Rang wird einem Wert I_{exp} sein normalisierter Wert I_{norm} aus der Urverteilung zugeordnet (roter Pfeil). Die blauen Linien geben die Schwankungsbreite eines exemplarischen Einzelwertes wieder ($\sigma = 32$). Die magentanen Pfeile visualisieren die Abbildung des Fehlers auf die Referenzverteilung.*

Referenzverteilung und Rangwertmittel

Nimmt man die Verteilungsähnlichkeit als absolutes Normalisierungskriterium, so muß es auch bei Realdaten eine Urverteilung geben. Hat man Kenntnis von dieser Urverteilung, so kann man jede abweichende Verteilung über die Ränge auf diese Referenzverteilung abbilden.

Leider ist man bei den Realdaten meist in Unkenntnis der wahren Urverteilung, daher muß man eine gute Näherung der Urverteilung finden (Referenzverteilung). Es kommen dazu mehrere Möglichkeiten in Betracht:

1. **Rangwertmittel** (Die mittlere Verteilung aller zu vergleichenden Experimente wird als Referenz herangezogen)

2. **Referenzexperiment** (Die Verteilung eines Referenzexperiment mit möglichst geringem Rauschen und optimalen Meßbedingungen dient als Referenz)

Wie oben beschrieben wird die Urverteilung durch Fehlereinflüße verändert. Die beste Näherung (zumindestens bei den Testdaten) ist das Experiment mit dem geringsten Fehler. Bei Realdaten sind aber nicht immer die Fehlereinflüsse bekannt und quantifiziert. In diesem Fall kann das Rangwertmittel eingesetzt werden. Ein Vergleich der beiden Möglichkeiten an Testdaten folgt weiter unten.

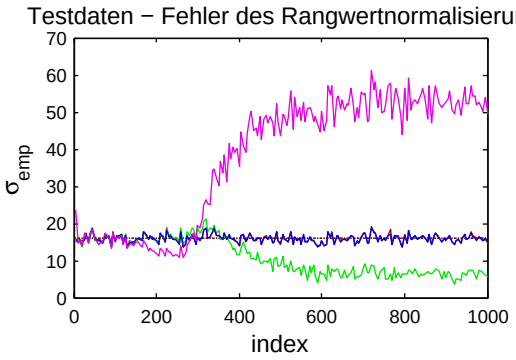


Abbildung 4.41: ***Vergleich der Normalisierung mit Referenzverteilungen*** *Hier ist die empirische Standardabweichung zwischen nicht normalisierten Testdaten (rote Kurve unterhalb der blauen) und normalisierten Testdaten (blaue Kurve - Rangwertmittel und grüne Kurve - Urverteilung als Referenz) dargestellt. Die magentane Kurve repräsentiert die Normalisierung auf ein Vergleichsexperiment mit höherem Rauschen ($\sigma = 32$). Die Abszisse ist der Index der Werte (entspricht dem Rang in der Urverteilung). Die schwarze Linie stellt die Standardabweichung des addierten Rauschens dar ($\sigma = 16$).*

Fehlerbetrachtung Der Rang gibt die Position eines Wertes innerhalb einer größengeordneten Reihe aller Werte an. Diese Position ist im wesentlichen eindeutig. Nur wenn mehrere gleichgroße Werte vorkommen wird diesem Wert ein Rangbereich zugeordnet. Für ein Gen hingegen ist die Zuordnung zu einem bestimmten Wert fehlerbehaftet. Demzufolge ist auch die Position innerhalb der Verteilungskurve für das Gen fehlerbehaftet. Die Schwankungsbreite des Wertes kann über die Standardabweichung definiert werden. Die Schwankungsbreite der Rangzuordnung kann nun dem Rangbereich zugeordnet werden, der zwischen oberer und unterer Fehlergrenze liegt.

Für das Rangwertmittel ergibt sich der Fehler des normierten Wertes aus folgender Überlegung. Das Zuordnen der Primärwerte zu den Normwerten erfolgt über den Rang. Der oben definierte Rangfehler gibt den Bereich an, innerhalb dessen die Rangzuordnung schwanken kann. Daher darf eigentlich nicht nur ein Einzelwert zugeordnet werden, sondern es muß der Wertebereich der Normverteilung zugeordnet werden, der der Schwankungsbreite des Rangfehlers entspricht. Abbildung 4.40 verdeutlicht dieses Konzept.

Ein weiterer zu beachtender Fehlereinfluß wäre der Rangwertfehler der Referenzverteilung. Dieser Fehler ist in Realdaten ohne Kenntnis der Urverteilung nicht zu bestimmen. Für die Testdaten ist es ohne Belang, da die fehlerfreie Urverteilung herangezogen werden kann.

Vergleich Für den Vergleich zwischen Rangwertmittel und Referenzverteilung werden Testdaten verwendet. Dabei werden 100 Experimente mit einem gleichgroßen Rauschen belegt ($\sigma = 16$). In Abbildung 4.41 wird diese Abweichung zwischen den Ausgangswerten mit der Standardabweichung der normalisierten Werte verglichen. Es zeigt sich, daß die Normalisierung mit dem Rangwertmittel (blaue Kurve) keine starken Einfluß auf den Fehler des normierten Einzelwertes hat. Dagegen zeigt die Normalisierung auf die Urverteilung eine Halbierung dieses Fehlers im unteren Wertebereich.

Wie kann diese Fehlerreduktion begründet werden? Die zusätzliche Normierungsinformation der Urverteilung führt die rauschveränderte Verteilung auf diese zurück. Das schließt, zum Beispiel in diesem Fall, negative Werte aus. Da jeder negative Wert auf einen positiven abgebildet wird, wird bei den unteren Werten der negative Einfluß des Rauschen reduziert.

Der gegenteilige Effekt wird erreicht, wenn ein Experiment mit höherem Rauschen ($\sigma = 32$ - magentane Kurve) als Referenz verwendet wird.

Die Normalisierung mit einem Referenzexperiment scheint für den niedrigen Wertebereich einen Vorteil

zu ergeben, wenn dieses Experiment näher an der Urverteilung liegt als die anderen zu normalisierenden Experimente. Hat die Referenz den gleichen Rauscheinfluß in ihrer Verteilung wie die zu normalisierenden Experimente, bleibt der Fehler konstant und es ergibt sich keine Verschlechterung dieses Fehlers. Die Referenz ist sehr ungünstig, wenn diese einen stärkeren Rauschanteil enthält als die zu normalisierenden Experimente.

Für das Rangwertmittel bedeutet das: Es sollte nur dort eingesetzt werden, wo man weiß, daß der Fehler der Experimentreihe ähnlich ist (gleiche Größenordnung). Es ist aber immer besser, ein Referenzexperiment mit einem sehr geringen Fehler einzusetzen.

Realdaten Für Realdaten geben die Ergebnisse aus den Testdaten nur einen Hinweis über den Einfluß der Normalisierung auf den relativen Fehler, da hier die Einflüsse auf die Verteilung vielfältiger sind als die gemachten Annahmen in den Testdaten. Was man übernehmen kann, ist die Bedeutung des Referenzexperimentes. Hier sollte ein sorgfältig ausgewähltes Experiment als Referenzverteilung verwendet werden. Das rRISA-Diagramm für die HAA1.2 Daten wird hier nicht gezeigt, da es nach der Normalisierung definitionsgemäß keine Abweichungen zwischen den Verteilungen der normalisierten Experimente gibt. Die rRISA-Kurve würde der Nulllinie entsprechen.

4.4.8 Fazit

Es zeigt sich deutlich, daß bei den linearen Methoden das asymmetrisch gestutzte Mittel bei geeigneter Parametrisierung die beste Wahl ist. Bei den Testdaten ergibt sich sogar eine leichte Verbesserung des relativen Normalisierungsfehlers $\frac{\overline{\Delta\kappa}}{\kappa}$ durch das obere Stutzen gegenüber der Mittelwertnormalisierung. Der Unterschied ist allerdings bei den Testdaten so gering, daß keine signifikante Verbesserung der Normalisierung (rRISA-Kriterium) erfolgt. Bei den HAA1.2-Realdaten kommt die Möglichkeit des Stutzens besser zum Tragen, da hier aufgrund von systematischen Meßfehlern (Sättigungseffekte) der hohe Meßwertbereich stärker fehlerbehaftet ist als der Rest der Verteilung. Hier verbessert ein 2%-Stutzen die Normalisierung sichtlich. Bei der Parameterwahl der Stutzung fällt auf, daß nur das obere Stutzen eine Verbesserung bringt. Unteres Stutzen bringen keinen Vorteil und hat bei Realdaten kaum einen Effekt. Daraus folgt, daß die Methode des symmetrischen Stutzen sich nahezu gleich der einseitigen oberen Stutzung verhält. Die anderen linearen Methoden bringen einen stärkeren Normalisierungsfehlers $\frac{\overline{\Delta\kappa}}{\kappa}$ ein. Hinsichtlich dieses Kriteriums ist die klassische Methode der Referenzgennormalisierung die schlechteste Wahl. Von den klassischen Methoden ist die Mittelwertnormalisierung die beste Wahl. Die Quantil- oder Perzentil-Normalisierung sind bei den Testdatensätzen generell schlechter als das globale Mittel.

Die Rangwertnormalisierung ist die einzige nichtlineare Methode die hier mitaufgeführt ist, da sie sich direkt aus den betrachteten rangbasierten Methoden (Perzentil, AGM) ableitet. Sie ist sehr mächtig bei der Anpassung der Verteilung aneinander. Bezüglich der Auswahl der Referenz konnte klar gezeigt werde, daß ein Referenzexperiment, das alle Qualitätskriterien erfüllt, am Besten dafür geeignet ist. Das Rangwertmittel stellt nur eine Alternative dar, wenn nicht entschieden werden kann, welches Experiment das qualitativ beste ist. Bei der Methode ist auffällig, daß sich die Standardabweichung zwischen verrauschten Testdaten signifikant durch die Normalisierung verändert. Je nach Referenzverteilung kann sie verstärkt oder verringert werden. Dieses Verhalten ist stark wertebereichsabhängig. Für Realdaten bedeutet das auch einen Einfluß auf die Signalstärke der wirklichen, biologisch generierten Veränderungen. Die Methode kann sehr gut experimentelle Veränderungen in der Meßfunktion ausgleichen. Allerdings werden auch lokale Veränderungen der Werteverteilung angepaßt. Somit besteht die Gefahr des „Overfitting“, da auch biologische Einflüße auf die Werteverteilung nihiliert werden.

Normalisierungsstrategie Die hier dargestellten Methoden haben verschiedene Vor- und Nachteile. Wie kann nun eine optimierte Normalisierungsstrategie aussehen? Generell sind Globalisierungsmethoden nur für größere Arrays (>500 Gene/Spots) geeignet. Nach einer Qualitätsbewertung erfolgt eine

Vornormalisierung mittels der Mittelwertnormalisierung. Damit sollten die Rang-Intensitäts-Kurven innerhalb eines RID vergleichbar sein. Wenn die Kurven visuell ähnlich erscheinen, folgt anschließend eine Bewertung mittels der relativen Rang-Intensitäts-Standardabweichung (rRISA) der mittelwertnormalisierten Experimente. Unterscheiden sich die Kurven dagegen stark voneinander, so sind Globalisierungsmethoden ungeeignet. Sind nur einzelne Kurven betroffen, so sollten die zugehörigen Experimente ausgeschlossen werden. Wenn man von einem konstanten statistischen Rauschen in den Daten ausgeht, so sollte die rRISA kontinuierlich im oberen Bereich zu den hohen Werten hin abnehmen. Ist dieses nach der Mittelwertnormalisierung der Fall, reicht diese bereits aus. Weicht nur ein kleiner Anteil der höchsten Werte (1-2%) von diesem Trend ab, so sollte der Bereich der Mittelwertbestimmung um diesen Bereich gestutzt werden. Das führt zu einer erneuten Normalisierung mittels des, hier eingeführten oberen gestutzten Mittels (OGM). (In ähnlicher Weise könnte ein extremer negativer Wertebereich, der nicht rauschbedingt verursacht wird, gestutzt werden: Wie oben gezeigt sind die Stutzungsparameter unabhängig).

Ist ein größerer Anteil der Werte von z.B. Sättigungseffekten beeinflußt, oder dieser Anteil beinhaltet sehr viele Meßpunkte (z.B. genomische Arrays von Affymetrix), so bietet sich die Verwendung einer nichtlinearen Methode an. Dafür kann die Rangwertnormalisierung benutzt werden. Hier ist zu entscheiden welche Referenz zu benutzen ist. Am günstigsten ist die Verwendung des qualitativ besten Experimentes (geringes Rauschen, lineare Meßfunktion, keine Sättigung zu erkennen).

Kapitel 5

Diskussion

Die vorliegende Arbeit besteht aus zwei Teilen. Der erste Teil beschäftigte sich mit der Entwicklung eines einfachen Hybridisierungsmodell zum Verständnis häufig auftauchender Fehler in Genexpressionsdaten. Das entwickelte Hybridisierungsmodell ist ein einfaches thermodynamisches Modell und basiert nicht direkt auf empirischen Daten von DNA-Arrays. Trotz dieser Limitation können Aussagen über das generelle Hybridisierungsverhalten auf Arrays getroffen werden. Insbesondere die Beschreibung der Kreuzhybridisierung (Abschnitt 4.1.6) kann Effekte erklären die man in Realdaten sieht. Der zweite Teil behandelt rangbasierte Methoden zur Bewertung, Fehlerabschätzung und Normalisierung von Genexpressionsexperimenten.

Hybridisierungsmodell In GEA-daten von Oligoarrays, speziell der Firma Affymetrix, werden Signale von Kreuzhybridisierungen durch gezieltes Sondendesign und den Algorithmen der mitgelieferten Auswertesoftware (Mikroarraysuite 5.0) ausgeschlossen, Dieses Vorgehen kann jedoch unter Umständen zu einer falschen Quantifizierung des wahren Expressionssignals führen, wie Chudin et al. ausführen [Chudin2001]. Die Autoren haben kontrollierte Mengen an Probenmoleküle (mRNA-fragmente bestimmter Moleküle) mit Affymetrix Hu95A-Arrays hybridisiert und ein unerwartetes Verhalten beschrieben. Über einen gewissen Mengenbereich verhält sich das gemessene Signal linear zur Probenmenge. Aber ab einer bestimmten Probenmenge nimmt das Signal wieder ab. Die Signalkurve ist somit nicht eineindeutig. Die Autoren führen diesen Effekt auf eine falsche Berechnungsfunktion in der damaligen Auswertesoftware von Affymetrix zurück und vermuten aber auch Kreuzhybridisierungsverhalten. Das letztendliche Expressionssignal wurde in jener Programmversion als die Differenz zwischen den Signalintensitäten der zugehörigen perfect-match-Sonden und mismatch-Sonden definiert. Die veraltete Berechnungsformel verstärkte nur einen Effekt, der ähnlich dem oben beschrieben ist. In Abbildung 5.1 sind zwei Diagramme dargestellt. Im linken Diagramm wurden die Daten aus dem Chudin-Artikel für das prokaryotische Kontrollgen ThrX_5 verwendet. Der Effekt in der resultierenden Signalkurve (gelb) ist leicht zu erkennen. Im rechten Diagramm ist ein ähnliches Verhalten zu beobachten. Hier entstanden die Kurven aus einigen geschätzten, Affymetrix typischen Parameter (siehe Tabelle 1.1) und den 25mer-BMP2 Sondenset (siehe Tabelle 4.1) unter Verwendung des Modell aus Abschnitt 4.1.5. Durch diese Ähnlichkeit läßt sich leicht auf die Ursache des eigentümlichen Signalverhaltens schließen. Bei den 25mer Sonden sind die Unterschiede in der Bindungsstärke zwischen Perfect-match und Mismatch nicht besonders groß. Dadurch kommt es bei höheren Probenmengen nicht nur zur Sättigung der Perfektmatchsonde mit den korrespondierenden Probenmolekülen, sondern auch gleichzeitig zu einer Kreuzhybridisierung mit der Mismatch-Sonde. Unter den gewählten experimentellen Bedingungen ist dieser Übergang nicht scharf, sondern abgerundet. Das Perfectmatch-Signal nähert sich mit steigender Probenmenge allmählich dem Grenzwert der Sättigung. Gleichzeitig steigt das Mismatch-Signal weiter an. Eine Differenzbildung führt nun zu dem oben beschriebenen Effekt. Das Mismatchsignal geht in den Chudin-Daten ungefähr bei der

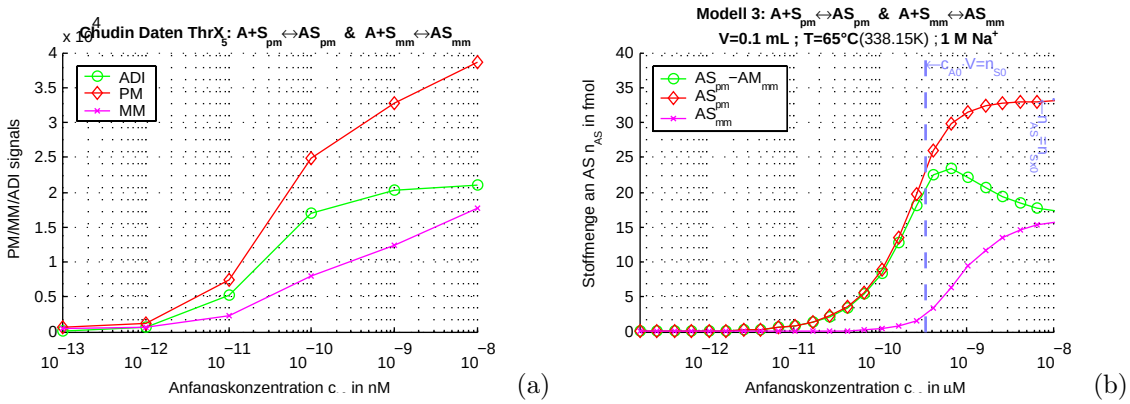


Abbildung 5.1: Vergleich von experimentellen Daten [Chudin2001] auf dem Hu95A-Array von Affymetrix (a) mit dem Modell 3 aus Abschnitt 4.1.5 dieser Arbeit (b). Die experimentellen Daten stammen von Testhybridisierungen von definierten Proben für das Kontrollgen ThrX_5 des Arrays. Die Parameter für das Modell mit den 25mer BMP2-Sonden (Tabelle 4.1) sind grob an die Daten angepaßt, sie stimmen wahrscheinlich nicht 100%ig mit dem Experiment überein. Aussagen dazu siehe Text.

Hälfte des Perfectmatchsignals in die Sättigung. Das läßt den Schluß zu, daß die molaren Mengen der verschiedenen Sonden nicht gleich sind. Es gibt ungefähr die doppelte Menge an Perfektmatchsonden als Mismatchsonden. Unter dieser Prämisse sind auch die Parameter für die Simulation gewählt. Es fehlen allerdings die Angaben zu den wirklichen Sondenmengen auf dem Chip, so daß unter Annahme der groben Firmenangaben aus der obigen Tabelle 1.1 auf Seite 9 nur eine grobe Übereinstimmung zwischen beiden Diagrammen erfolgen kann. Mit geeigneter Parameterwahl könnte die quantitative Übereinstimmung noch verbessert werden, aber dieses Fitten würde das einfache Modell überbewerten, da die Richtigkeit der angepaßten Parameter nicht überprüft werden kann. Andererseits sind die Chudindaten nicht dicht genug um auf ihnen ein Modell aufzubauen.

Dieses Beispiel zeigt sicherlich die Grenzen des Modells. Aber auch die Grenzen der GEA-Methode, da ohne Kenntnis der unbekannten Parameter nicht zwischen normaler und mengeninduzierter Kreuzhybridisierung unterschieden werden kann. Der Effekt ist ein Beispiel für schlecht normalisierbare Einflüße. Überhaupt sind noch viele Kooperativeffekte auf dem Array zu wenig untersucht. Die Kreuzhybridisierung zwischen Probenmolekülen ist so ein Fall. Die Simulation mit sehr überschaubaren Grundannahmen (Abschnitt 4.1.9) zeigt eine Wurzelfunktion als Hybridisierungsfunktion. Der Nachweis dieser Nebenreaktionen in Daten von normalen GEA-Experimenten ist kaum möglich, da die wirklichen Stoffmengen der einzelnen Probenmolekülen unbekannt sind. Werden alle Proben mit dem selben Protokoll verarbeitet, kann ohne zusätzliche Information an Hand der Daten nicht entscheiden werden, ob die Hybridisierungsfunktion linear ist oder es sich um eine Wurzelfunktion handelt. Die allgemein verbreitete Grundannahme in der GEA-Analyse ist die Linearität der Hybridisierungsfunktion für die meisten Proben, Das scheint eine funktionierende Arbeitshypothese zu sein. Man muß aber auch, wie beschrieben, nichtlineare Abweichungen in Betracht ziehen.

Zusammengefaßt kann man sagen, daß das entwickelte Modell eine gute Grundlage bildet, unerwartete Effekte in Experimentaldaten zu erklären und zu behandeln. Viele der Effekte sind aufgrund der fehlenden Information mit mathematischen Mitteln nicht zu korrigieren. Im Einzelfall muß man versuchen den auftretenden Fehler durch Optimierung der Bedingungen (z.B. Veränderungen der Probenmenge) zu minimieren. Für quantitative Aussagen muß das Modell allerdings noch wesentlich erweitert werden, speziell um eine kinetische Komponente und um die Berücksichtigung der Diffusion.

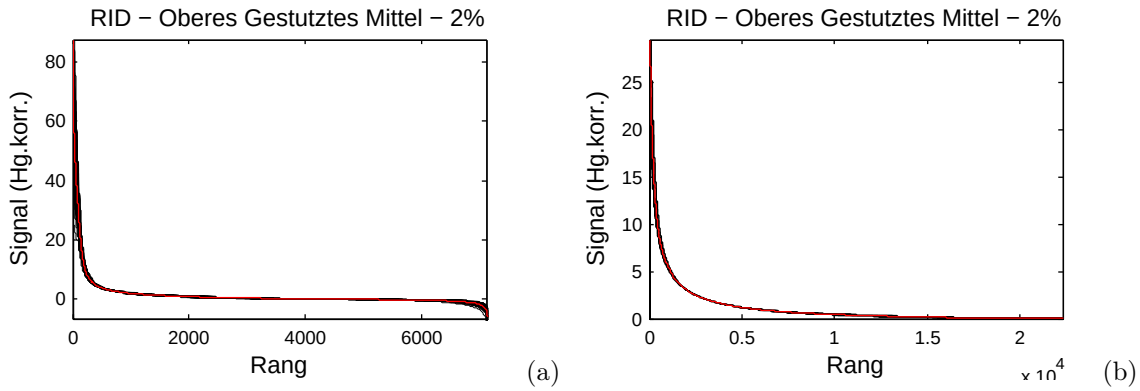


Abbildung 5.2: Rang-Intensitäts-Kurven normalisiert auf den ***2%-oberes gestutztes Mittel*** *des Einzelexperiments. Die Daten sind von Experimenten von AffymetrixArrays (a) Hu95A aus Literaturdaten [Golub1999] und (b) Hu133A von Experimenten mit Zellkulturen von K562-Zellen (rote Kurve ist das Rangmittel)*

Visualisierung und Werteverteilung Um eben diese Probleme in den Daten sichtbar zu machen, bedarf es einer guten Visualisierung. Herkömmliche Scatterplots zwischen zwei Experimenten reichen für eine Qualitätsbegutachtung meist nicht aus. Besonders für die gleichzeitige Bewertung mehrerer Experimente bedarf es anderer Darstellungsformen. Dabei erwiesen sich für die vorliegende Arbeit rangbasierte Diagramme als besonders hilfreich, globale Dateneigenschaften zu erfassen. Die Form des Rang-Intensitäts-Diagramms (RID) wurde bis dato [Kroll2002B] nicht auf Genexpressionsdaten angewandt. Dabei ist dieses Diagramm besser geeignet, die Werteverteilungen zu vergleichen als ein Histogramm. Das RID zeigt beispielsweise Sättigungseffekte als ein Plateau der Rangwertfunktion im Bereich der hohen Werte an. Im Histogramm ist dieses Verhalten schwerer zu erkennen, da es sich um einem wenig besetzten Wertebereich handelt. Andererseits stellt das RID nur globale Eigenschaften der Daten dar, da der Genbezug aufgehoben ist und sich lokale Arrayeffekte in der Verteilung verstecken können. Besonders für die Bewertung der Normalisierung stellt das aber keinen Nachteil dar, da die Normalisierung eine globale Optimierung der experimentellen Vergleichbarkeit darstellt.

Das Rang-Intensitäts-Diagramm ist wie in Abschnitt 4.4 gezeigt, gut geeignet die Werteverteilung von mehreren Experimenten miteinander zuvergleichen. Das Diagramm ist an zwei Datensätzen eingeführt worden. Einer davon ist ein Testdatensatz. In wieweit sind die getroffenen Aussagen auch für andere Arraydaten gültig?

Generell ist die Form der Verteilung von der Art der Probe und dem ausgewählten konkreten Arraytyp abhängig. Vergleicht man nur Experimente eines Arraytyps miteinander scheinen die Rang-Intensitäts-Kurve bei allen von mir bearbeiteten Arrays ähnlich zu sein. Diese Ausage ist allerdings nicht prinzipiell verallgemeinerbar. Ich erwarte z.B. für spezialisierte Arrays mit geringer Sondenanzahl durchaus extreme Unterschiede in der Werteverteilung, da diese vorwiegend eine Auswahl extrem regulierte Gene enthalten, für die eine notwendige Grundannahme der meisten Normaliserunsverfahren nicht gilt: Die Gesamtmenge an arraygebundener Probenmoleküle (mRNA oder cDNA) ist proportional zur Gesamtmenge an mRNA der Probe.

Abbildung 5.2 zeigt die RID verschiedener Arraytypen von normalisierten Daten (4% oberes gestutztes Mittel - OGM - siehe Abschnitt 4.4.6). Weichen einzelne Kurven, trotz prinzipieller biologischer Ähnlichkeit der verwendeten Proben, deutlich von der Grundgesamtheit der Rang-Intensitäts-Kurven ab, ist das, nach meiner Erfahrung, immer ein deutliches Zeichen für grobe experimentelle Fehler. Diese konnten meist auf Alterungserscheinungen mehrfach verwendeter Membranen zurückgeführt werden und waren bereits im Vorfeld durch schlechtere Bildeigenschaften aufgefallen.

(Ein direkter Vergleich der Rang-Intensitäts-Kurve verschiedener Arrays ist bei unterschiedlicher Sondenanzahl (Ränge) mit dem einfachen RID nicht möglich. Es besteht aber die Möglichkeit der Rangnormierung, d.h. es werden nicht die Ränge selbst aufgetragen, sondern der relative Rang (Rangzahl/Gesamtranganzahl).)

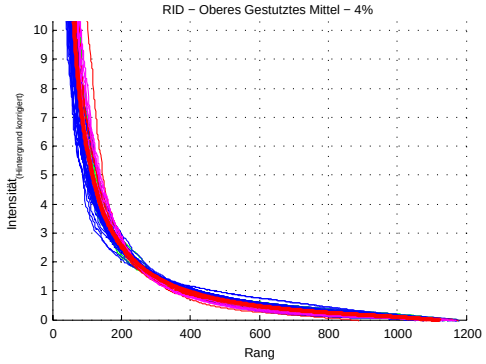


Abbildung 5.3: Rang-Intensitäts-Kurven normalisiert auf den ***4%-oberes gestutztes Mittel*** *des Einzelexperiments. Die Daten sind von HAA1.2 Arrays. Erklärung siehe Text. (rot und lila ↦ Blutproben / grün und blau ↦ Nieren- und Lungenproben)*

In Abbildung 5.3 ist eine Auswahl der verwendeten Evaluierungsdaten der Clontechmembranen (HAA1.2) zusätzlich farbgruppiert, d.h. Experimente einer Probengruppe haben die gleiche Farbe. Alle Daten sind auch hier mit 4% OGM normalisiert. Es zeigen sich leichte Unterschiede zwischen den Gruppen. Zwei Tendenzen werden deutlich die roten/lilanen Linien (R/L) gruppieren zusammen und die blau/grünen (B/G). Die R/L haben eine steilere Werteverteilung als die B/G, das bedeutet, daß der Unterschied zwischen den niedrig und hoch exprimierten Werten bemerkbar größer ist. Da die dazugehörigen Experimente nebeneindander im gleichen Zeitraum ausgeführt wurden, liegt der sichtlichen Eigenschaftsgruppierung wahrscheinlich keine Meßparameterunterschiede zu Grunde, sondern ein systematischer Probenunterschied. Bezüglich der beiden ausgewählten Probengruppen ist eine Erklärung offensichtlich. Bei den R/L Linien handelt es sich um Blutproben von Leukämiepatienten zweier unterschiedlicher Experimentalreihen (siehe M&M). Die B/G Gruppe besteht aus Daten von Lungen- und Nierengewebe. Die Blutproben stellen eine viel homogenere Zellgruppe dar als die Gewebeproben, die viele unterschiedliche Zelltypen (teilweise tumorhaltig) in sich vereinen. Jeder Zelltyp hat sein eigenes typisches Expressionsmuster. Mischt man diese, kommt es (auch ohne Kooperativeffekte) zu einem gemischten Muster. Viele Expressionswerte von Genen, die nur in einem Zelltyp hochexprimiert sind, sind in der Mischung verringert. In der Werteverteilung wird der Bereich der mittelmäßig exprimierten Gene vergrößert. Der Abfall der Rang-Intensitäts-Kurve wird umso flacher je verschiedener die Zellen in der gemessenen Probe sind. Die Erklärung der Verteilungsunterschiede ist also klar auf prinzipielle Probenunterschiede zurückzuführen. (Das ist aber nicht nur als zusätzliches Problem zu verstehen. Es gibt auch die Bestrebung, innerhalb statistischer Grenzen, die Zusammensetzung einer Mischpopulation aus den Expressionsdaten der Mischung und der in ihr enthaltenen diskreten Zelltypen zu bestimmen. Was z.B. wichtig ist, um Veränderungen in der Gesamtgenexpression der Mischung von der puren Veränderung der Zusammensetzungsverhältnisse zu unterscheiden. [Hoffmann2002])

Normalisierung Bezüglich der Normalisierung heißt das, je ähnlicher die zu vergleichenden Proben biologisch sind, desto ähnlicher ist ihr Werteverhalten und desto besser lassen sich die eingeführten Kri-

terien, wie z.B. die relative Rang-Intensitäts-Standardabweichung (rRISA), anwenden. Es zeigt sich dadurch deutlich, daß die Grundannahmen, die für die diversen Normalisierungsmethoden gemacht werden, auch immer auf ihre Gültigkeit bei den konkreten Datensätzen überprüft werden müssen. Andererseits ist das ein Hinweis darauf, daß es bisher, wie in der Systemanalyse beschrieben, keine eindeutige Signalfunktion für die gesamte GEA-Methode gibt. Da es nun verschiedene unbestimmte Meßparameter gibt, hängt die Vergleichbarkeit der Ergebnisse im Wesentlichen von einer möglichst hohen Konstanz dieser Parameter ab. Das ist um so besser möglich, je ähnlicher das Meßsystem ist.

Der *status quo* der GEDA ist, die Signalfunktion wird als weitgehend linear angenommen und jede Abweichung wird mittels einer geeigneten Normalisierungsmethode linearisiert. Ist das nicht möglich, weil die nichtlinearen Einflüße keine Rekonstruktion zulassen (z.B. Signalbeschneidung durch oberes Detektionslimit beim Fluoreszenzscannen), läßt sich zu mindestens der lineare Bereich bestimmen. In beiden Fällen lassen sich dadurch unterschiedlich skalierte Experimente mittels einer linearen Normalisierung vergleichbar machen. Jedoch wirft die Bestimmung des Skalierungsfaktors eine biologische Frage auf: Wenn keine quantitativen Aussagen getroffen werden können, braucht man immer eine Referenz, auf die man sich bezieht, um relative, semiquantitative Aussagen treffen zu können. Welche Referenz ist biologisch sinnvoll?

Die Frage ist ohne genaue Kenntnis des einzelnen vermessenen biologischen Systems nicht zu beantworten. Alle bisher erwähnten möglichen Referenzmaße (konstante mRNA-Menge, konstante Expression eines oder mehrerer Referenzgene) sind unter Umständen biologisch falsch. Wenn sie gleichberechtigt sind, sollte dann die Referenz und die auf ihr basierende Normalisierungsmethode verwendet werden, die den geringsten mathematischen Normalisierungsfehler in die Daten einbringt.

In der vorliegenden Arbeit sind verschiedene, vorwiegend rangbasierte Skalierungsmethoden unter diesem Aspekt verglichen worden. Dafür wurden Testdaten benutzt, um eine genaue Kontrolle der Einflußgrößen zu haben. Als Modell der Werteverteilung von Realdaten wurde eine exponentielle Verteilung benutzt. Diese Annahme ist nur eine grobe Näherung. Realverteilungen von GEA-Daten sind immer abhängig von der Genauswahl auf dem Array. Die in der Arbeit benutzten HAA1.2-Daten liefern überexponentielle Verteilungen. Trotzdem werden durch die Näherung die wesentlichen Merkmale der meisten Realverteilungen simuliert: ein großer Bereich niedrig exprimierter Werte und ein kleiner Bereich hochexprimierter.

Auf die mit einem normalverteilten Rauschen versetzten Testdaten wurden nun verschiedene Normalisierungen angewendet. Als Fehlerkriterium wurde der relative Fehler des Skalierungswertes ausgewählt. Er stellt die direkte Fehlerverstärkung des relativen Fehlers auf des einzelnen normalisierten Meßwertes dar. Für die Testdaten konnte nun die Normalisierung ausgewählt werden, die diesen Fehler minimiert. Allerdings wurde im Abschnitt 4.3 auf Seite 53 gezeigt, daß die Urverteilung systematisch durch das Rauschen verändert wird. Daraus folgt für rangbasierte Methoden ein zweites wichtiges Kriterium. Die relative Standardabweichung der verrauschten Verteilung von der rauschfreien Urverteilung.

Für die Testdaten zeigt sich, daß Referenzmethoden, die auf einzelnen „house keeping"-Genen beruhen, den größten rauschinduzierten Normalisierungsfehler einbringen. Dabei gilt, je größer die Expressionsstärke des Gens desto geringer der eingeführte relative Fehler (bei gleichen absoluten Meßfehler). Den geringsten Fehler bringt die Benutzung des globalen Mittelwertes ein. Bei rangbasierten Methoden ergibt sich folgendes Bild: Perzentile bzw. Quantile ergeben generell einen größeren Normalisierungsfehler als der Mittelwert. Höhere Perzentile sind dabei besser geeignet als niedrigere. Durch die Verwendung des Urverteilungskriteriums zeigt sich noch eine interessante Eigenschaft. Der systematische Einfluß des Rauschens auf die Werteverteilung reicht noch bis hoch in die mittleren Perzentile ($<75\%$). Überhaupt geeignet waren nur Perzentile in der Nähe des Maximalwertes ($<100\%$ Perzentil). Die andere Gruppe rangbasierter Normalisierungen, die gestutzten Mittel, verhält sich bei optimaler Parametrisierung günstiger. Es zeigt sich, daß die klassische Methode des symmetrischen gestutzten Mittels keinen Verbesserung der Normalisierung gegenüber dem globalen Mittelwert bringt. Nur eine vollständige Stutzung um den rauschbeeinflußten niedrigen Rangbereich durch das untere gestutzte Mittel bringt eine bessere Normalisierung. Bei konstantem wertebereichsunabhängigen Absolutfehler bringt dagegen die oberer

Stutzung keinen Vorteil gegenüber dem globalen Mittelwert.

Für die Realdaten von Experimenten mit dem HAA1.2 Filter von Clontech kann leider nicht das gleiche Kriterium genommen werden, da bei wirklichen Experimenten die Urverteilungen (die wahren Werte) unbekannt sind. Hier kann das rRISA-Kriterium zum Einsatz kommen. Die relative Rang-Intensitäts-Standardabweichung ist ein Maß für die relative Ähnlichkeit der Werteverteilungen. Je besser also die Kurven aneinander durch die Normalisierung angenähert werden, desto besser ist die Grundannahme der rangbasierten Normalisierungsmethoden erfüllt: die Gleichheit der Werteverteilungen eines Arrays. Um das beurteilen zu können, ist eine Vornormalisierung notwendig. Aus der Betrachtung der Testdaten heraus ergibt sich die Benutzung des globalen Mittels als primär beste Methode.

Das rRISA-Diagramms der mittelwertnormalisierten Werte zeigt einen Bereich an, der dem Trend der Verringerung der relativen Abweichung der Werteverteilungen mit steigendem Wert entgegensteht. Dieser Bereich hat wahrscheinlich durch Sättigungseffekte oder extremregulierte Einzelgene eine höhere Rangwertabweichung. Dieser höhere Fehler geht auch in die Bestimmung des Skalierungsquotienten ein. Hier bietet sich eine Stutzung um diesen Bereich an, was zu Verwendung des oberen gestutzten Mittels (OGM) führt. Bei Testdaten zeigt sich eine generelle Verschlechterung des Normalisierungsfehlers durch das obere Stutzen. Es sollte also so wenig wie möglich gestutzt werden. Eine 1-2% obere Stutzung reicht aus, um die rRISA im Bereich der hohen bis mittleren Werte zu verringern. Eine Verringerung der rRISA im untersten Wertebereich ist dagegen nicht anzustreben, da hier systematische Veränderungen der ursprünglichen Werteverteilung durch das statistische Rauschen auftreten. Eine rRISA-Optimierung in diesem Bereich, würde zu einer Normalisierung auf eben diesen Fehlereinfluß führen. Eine untere Stutzung hat bei diesen Daten einen sehr geringen Einfluß. Sie ist daher nicht notwendig. Das führt dazu, daß eine geringe symmetrische Stutzung des Mittelwertes einen ähnlichen Effekt hat wie die entsprechende obere Stutzung. Es zeigt sich also, daß die Normalisierung mit einem 1-2% oberen gestuzten Mittel (OGM) die beste Normalisierung für den vorliegenden Datensatz bringt. Das gilt natürlich nur solange, wie auch die Grundannahmen (s.o.) dieser Normalisierung gültig sind [Kroll2002B].

Natürlich stellt sich die Frage, warum für diesen Datensatz nicht auch nichtlineare Normalisierungen eingesetzt wurden. Nichtlineare adaptive Normalisierungen, wie die vorgestellte Rangwertnormalisierung oder die Methode der lokale Regression, bedürfen eines Referenzexperimentes. An dessen Werteverteilung wird das zu normalisierende Experiment angepasst. Der vermutete nichtlineare Bereich umfasst etwa die 20-40 höchsten Werte mit einem starken zufälligen Fehler. Wie bei der Betrachtung des Perzentils/Quantils gezeigt, wird der Fehler durch die Rangordnung in diesem Wertbereich nicht verringert. In die Normalisierung geht also direkt der Fehler des Einzelwertes ein. Es ist sehr wahrscheinlich, daß die nichtlineare Normalisierung hier einen stärkeren Fehler einführt als die obige Skalierung.

Mit Datensätzen von größeren Arrays muß diese Einschätzung dagegen nicht zutreffen (z.B. Affymetrix-Arrays oder genomische cDNA-Glasarrays). Umfasst der nichtlineare Bereich eine höhere Anzahl von Werten und ist dieser Bereich kontinuierlich, läßt sich ein systematischer funktioneller Zusammenhang vermuten. Der Einsatz nichtlinearer Normalisierungen ist dann gerechtfertigt. Meist kann dieser nicht physikalisch funktionell erfasst werden. In diesen Fällen werden adaptive Methoden verwendet (Lowess- lokale Regression bei cDNA-Glasarrays [Yang2002A]). Die Gefahr hierbei liegt in einer physikalisch nicht gerechtfertigten Transformation der Werte. Das zeigt sich z.B. in der Fehlerbetrachtung der Rangwertmethode.

Die Rangwertmethode ergibt sich aus der Grundannahme der rangbasierten Normalisierungen: Die Gleichheit der Rang-Intensitäts-Verteilungen. Ein direktes Zuordnen der Ranges eines Genes zu einer Referenzverteilung liefert den dazugehörigen Referenzrangwert als normalisierten Wert zurück. Eine anderer Zugang zu dieser Methode läuft über laufende Quantile. Da die Quantile direkt mit den normierten Rängen verknüpft sind, kann man von äquivalenten Methoden ausgehen. Daher wird diese Normalisierung auch Quantilnormalisierung[1] genannt. Sie wird schon seit längerer Zeit diskutiert aber erst 2003 von Bolstad et al. [Bolstad2003] in einem Vergleich mit der Lowess-Methode publiziert. Laut Bolstad

[1] nicht zu verwechseln mit der Skalierung mit einzelnen Quantilen/Perzentilen

et al. bringt sie bei Affymetrix-Arrays gegenüber Lowess einen geringen Vorteil. Vor allem beim paarweisen Vergleich schafft sie die beste Anpassung. Die in dieser Arbeit durchgeführte Fehleranalyse an Testdaten zeigt allerdings, daß die optimale Anpassung zwischen zwei Experimenten nicht unbedingt zu Verringerung des Meßfehlers führt. In Abhängigkeit vom Fehler des Referenzexperimentes kann die Normalisierung den Meßfehler der niegrig exprimierten Gene verstärken oder verringern. Nur die Verwendung des am wenigsten fehlerbehafteten Experimentes als Referenz bringt einen klaren Vorteil. Die Analyse zeigt daher auch die Notwendigkeit der Bestimmung des Fehlers des Einzelexperimentes.

Eine Strategie dazu liefert diese Arbeit in Abschnitt 4.3. Allerdings wiederum „nur“ für HAA1.2-Daten und an Hand von Testdaten wird gezeigt, wie der Anteil des statistischen Meßrauschens retrospektiv in diesen Daten abgeschätzt werden kann. Notwendig dafür ist das Auftreten negativer Werte in den Daten. Es ließ sich ein vom Autor vermuteter Zusammenhang des Mittelwertes der negativen Werte mit der Standardabweichung des Rauschens bestätigen. Der für die Testdaten gefundene Wert des Proportionalitätsfaktors von rund 1.2 liegt sehr nahe am theoretischen Wert für die Nullverteilung von $\sqrt{\frac{\pi}{2}}$. Da die empirische Werteverteilung mit höheren Rängen noch steiler gegen die Nullwerte fällt als die Exponentialverteilung der verwendeten Testdaten sollte sie sich noch besser an diesen theoretischen Wert annähern. Allerdings sind in den realen Daten nicht nur statistisches Rauschen enthalten, sondern auch andere systematische Fehler, so daß der mittels der vorgestellten Methode geschätzte Meßfehler für einzelne Werte stark vom wahren Fehler abweichen kann. In wieweit diese Methode auf andere Daten übertragen werden kann, hängt stark vom jeweiligen Meßsystem ab. Hier könnten noch weitere Untersuchungen folgen.

Abschließend zur Normalisierung muß gesagt werden, daß dieses Gebiet sich sehr schnell weiterentwickelt. Die Methoden passen sich immer mehr an die jeweiligen Meßsysteme und ihrer Fehler an. Gestutzte Mittel werden wahrscheinlich nur noch für kleinere Arrays ($\approx$ 1000 Gene) Anwendung finden und als Methoden der Vornormalisierung [Affymetrix2002MAS]. Auch die adaptiven nichtlinearen Normalisierungen werden in Zukunft durch die Weiterentwicklung und qualitativen Verbesserung der Meßtechnik zurückgedrängt werden. Generell ist die Tendenz zu einer Standardisierung der einzelnen Meßschritte und konsequenten Anwendung von mitgeführten Kontrollen zu begrüßen. Peppel et al. [Peppel2003] konnte z.B. zeigen, daß eine retrospektive Methode wie Lowess deutlich ein anderes Normalisierungsmuster erzeugt als die externe Normalisierung mittels Spikes (zugemischte probenunabhängige Kontrollen). Besonders lokale Veränderungen zwischen Experimenten wurden deutlich durch Lowess abgeschwächt. Zumindestens diese Gefahr der lokalen Überanpassung besteht bei Skalierungsmethoden nicht.

Kapitel 6

Zusammenfassung

Die Genexpressionsmessung mit DNA-Arrays ist eine sehr komplexe und dadurch fehleranfällige Methode. Jeder der notwendigen Einzelschritte der Messung beeinflußt das letztendliche Meßergebnis durch verschiedene Parameter in signifikanter Weise. Im Gegensatz zu herkömmlichen statistischen Daten mit vielen Messungen und wenig Meßparametern beinhalten die Arraydaten sehr viele Meßparameter und nur wenig Messungen. Der Normalisierung kommt damit eine zentrale Rolle in der Datenanalyse zu. Diese soll die Meßparameter berücksichtigen und die Meßdaten auf die zumessende Größe „Anzahl einer bestimmten mRNA-Spezies in einer definierten Probe“ zurückführen oder zumindestens eine weitgehende Vergleichbarkeit zwischen Experimenten herstellen.

Im Rahmen dieser Arbeit wurde ein einfaches Hybridisierungsmodell entwickelt, um Einflüsse von Meßparametern auf die Signalfunktion abzuschätzen. Durch das Hybridisierungsmodell kann der zentrale Meßschritt der Hybridisierung der immoblisierten DNA-Sonde mit der in Lösung befindlichen cDNA-Probe simuliert werden. Hiermit können die Art und Stärke der Haupteinflußgrößen abgeschätzt werden. Für quantitative Aussagen sind die dem Modell zugrunde liegenden Annahmen zu einfach und die verfügbaren Informationen zu unvollständig. Qualitative Vorhersagen sind dagegen möglich. So kann z.B. mengeninduzierte Kreuzhybridisierung auf Oligoarrays beschrieben und verstanden werden.

Der zweite und Hauptteil dieser Arbeit beschäftigt sich mit der Normalisierung von cDNA-Filterarrays. Dazu wurde mehrere Skalierungsmethoden und eine nichtlineare adaptive Methode miteinander verglichen. Dabei konnte gezeigt werden, daß die vom Autor entwickelte Normalisierungsmethode des asymmetrisch gestutzten Mittels bei geeigneter Parametrisierung in Bezug auf Testdaten den geringsten Normalisierungsfehler verursacht. Für die Realdaten stellte sich ein leichter Vorteil der Normalisierung durch den Spezialfall des 2%igen oberen gestutzten Mittels heraus, da es damit möglich ist den Einfluß von Sättigungseffekten etc. auf den Skalierungsquotienten zu minimieren. Ausführlich wurden dabei die Grundlagen der rangbasierten Normalisierungsmethoden erörtert und der Fehlereinfluß der nichtlinearen Rangwertnormalisierung diskutiert. Für die Qualitätsbeurteilung und Normalisierungskontrolle erwiesen sich zwei rangbasierte Kriterien als sehr nützlich: die Rang-Intensitäts-Kurven und die davon abgeleitete relative Rang-

Intensitäts-Standardabweichung. Beide Kriterien wurden dabei erstmals für die Genexpressionsanalyse angewandt. Ein weiterer Aspekt der zugehörigen Visualisierungen lieferte die Grundlage für die retrospektive Fehlerabschätzung aus realen Genexpressionsdaten. Es wurde gezeigt, daß die mittleren negativen Werte der Rang-Intensitäts-Verteilungen sich proportional zur Standardabweichung des statistischen Rauschanteils verhalten.

Kapitel 7

Literaturverzeichnis

[1] Quellenangaben zu zitierten Artikeln und Produkthandbüchern

[Adams1993] M.D. Adams, A.R. Kerlavage und C. Fields: „*3,400 new expressed sequence tags identify diversity of transcripts in human brain*“ **Nature Genetics** (1993), Band 4, Nummer 3, Seiten 256-267

[Adams1995] M.D. Adams, A.R. Kerlavage, R.D. Fleischmann, R.A. Fuldner, J.D. Gocayne und O. White: „*Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence*“ **Nature** (1995), Band 377, Seiten 3-174

[Adorjan2002] P. Adorjan, J. Distler, E. Lipscher, F. Model, D. Gütig, G. Grabs, A. Howe, M. Kursar, R. Lesche, E. Leu, A. Lewin, S. Maier, V. Müller, T. Otto, C. Scholz,et al. :„*Tumour class prediction and discovery by microarray-based DNA methylation analysis*“ **Nucleic Acids Research** (2002), Band 30, Nummer 5, Seiten e21

[Affymetrix2001] Firma Affymetrix: „*25mer Sensitivity and specificity*“ **Affymetrix TechNote** (2001), Nummer 701009

[Affymetrix2002MAS] Firma Affymetrix: „“ **Affymetrix Microarray Suite User Guide. Version 5** (2002)

[Affymetrix2002S] Firma Affymetrix: „*GeneChip® Expression Analysis Data Analysis Fundamentals*“ **Affymetrix TechNote**

[Alizadeh2000] A.A. Alizadeh, M.B. Eisen, E.R. Davis, C. Ma, T. Tran, X. Yu, J.I. Powel, L. Yang, G.E. Marti, T. Moore, J.J.r. Hudson, L. Lu, D.B. Lewis, R. Tibshirani, G. Sherlock,et al. :„*Distinct types of diffuse large B-cell lymphoma identifed by gene expression profling*“ **Nature** (2000), Band 403, Seiten 503-512

[Allawi1997] H.T. Allawi: „*Thermodynamics and NMR of Internal G,T Mismatches in DNA*“ **Biochemistry** (1997), Band 36, Seiten 10581-10594

[Allawi1998A] H.T. Allawi: „*Nearest Neighbor Thermodynamic Parameters for Internal G,A Mismatches in DNA*“ **Biochemistry** (1998), Band 37, Seiten 2170-2179

[Allawi1998B] H.T. Allawi: „*Nearest-Neighbor Thermodynamics of Internal A,C Mismatches in DNA: Sequence Dependence and pH Effects*“ **Biochemistry** (1998), Band 37, Seiten 9435-9444

[Allawi1998C] H.T. Allawi: „*Thermodynamics of internal C·T mismatches in DNA*“ **Nucleic Acids Research** (1998), Band 26, Nummer 11, Seiten 2694-2701

[Beissbarth2000] T. Beissbarth, K. Fellenberg, B. Brors, R. Arribas-Prat und A. Poustka: „*Processing and quality control of DNA array hybridization data*“ **Bioinformatics** (2000), Band 16, Nummer 11, Seiten 1014-1022

[Berns2000] A. Berns: „*Gene expression in diagnosis*“ **Nature** (2000), Band 403, Seiten 491-492

[Bertucci1999] F. Bertucci, K. Bernard, B. Loriod und Y.C. Chang: „*Sensitivity issues in DNA array-based expression measurements and performance of nylon microarrays for small samples*“ **Human Molecular Genetics** (1999), Band 8, Nummer 9, Seiten 1715-1722

[Bhattacharjee2001] A. Bhattacharjee, W.G. Richards, J. Staunton, C. Li, R. Bueno, M. Gillette, M. Loda, G. Weber, E.J. Mark, E.S. Lander, W. Wong, B.E. Johnson, T.R. Golub und D.J. Sugarbaker: et al. :„*Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses*“ **Proceedings of the National Acadamy of Science** (2001), Band 98, Nummer 24, Seiten 13790-13795

[Bittner2000] M. Bittner, P. Meltzer, Y. Chen, Y. Jiang, Z. Yakhinik, A. Ben-Dork, N. Sampask, E. Dougherty, E. Wang, F. Marincola, C. Gooden, J. Lueders, A. Glatfelter, P. Pollock, J. Carpten,et al. :„*Molecular classification of cutaneous malignant melanoma by gene expression profiling*“ **Nature** (2000), Band 406, Seiten 6795-6799

[Black2002] M.A. Black: „*Calculation of the minimum number of replicate spots required for detection of significant gene expression fold change in microarray experiments*“ **Bioinformatics** (2002), Band 18, Nummer 12, Seiten 1609-1616

[Blake1999] R.D. Blake, J.W. Bizzaro, J.D. Blake und G.R. Day: „*Statistical mechanical simulation of polymeric DNA melting with MELTSIM*“ **Bioinformatics** (1999), Band 15, Nummer 5, Seiten 370-375

[Bockelmann2002] U. Bockelmann, P. Thomen, B. Essevaz-Roulet und V. Viasnoff: „*Unzipping DNA with optical tweezers: high sequence sensitivity and force flips*“ **Biophysical Journal** (2002), Band 82, Seiten 1537-1553

[Bolstad2003] B.M. Bolstad, R.A. Irizarry, M. Astrand und T.P. Speed: „*A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.*“ **Bioinformatics** (2003), Band 19, Nummer 2, Seiten 185-93

[Bommarito2000] S. Bommarito und N. Peyret: „*Thermodynamic parameters for DNA sequences with dangling ends*“ **Nucleic Acids Research** (2000), Band 28, Nummer 9, Seiten 1929-1934

[Bonnet1998] G. Bonnet und O. Krichevsky: „*Kinetics of conformational fluctuations in DNA hairpin-loops*“ **Proceedings of the National Acadamy of Science** (1998), Band 95, Seiten 8602-8606

[Bonnet1999] G. Bonnet, S. Tyagi und A. Libchaber: „*Thermodynamic basis of the enhanced specificity of structured DNA probes*“ **Proceedings of the National Acadamy of Science** (1999), Band 96, Seiten 6171-6176

[Breslauer1986] K.J. Breslauer, R. Frank und H. Blocker: „*Predicting DNA duplex stability from the base sequence*“ **Proceedings of the National Acadamy of Science** (1986), Band 83, Nummer 11, Seiten 3746-3750

[Bustin2000] S.A. Bustin: „*Absolute quantification of mRNA using real-time reverse transcription polymerase chain reaction assays*“ **Journal of Molecular Endocrinology** (2000), Band 25, Seiten 169-193

[Bustin2002] S.A. Bustin: „*Quantification of mRNA using real-time reverse transcription PCR (RT-PCR): trends and problems*“ **Journal of Molecular Endocrinology** (2002), Band 29, Nummer 1, Seiten 23-39

[Causo2000] M.S. Causo und B. Coluzzi: „*Simple model for the DNA denaturation transition*“ **Physical Reviews E** (2000), Band 62, Nummer 3, Seiten 3958-73

[Chalikian1999] T.V. Chalikian, J. Volker und G.E. Plum: „*A more unified picture for the thermodynamics of nucleic acid duplex melting: a characterization by calorimetric and volumetric techniques*“ **Proceedings of the National Acadamy of Science** (1999), Band 96, Seiten 7853-7858

[Chen1997] Y. Chen, E.R. Dougherty und M.L. Bittner: „*Ratio-based decisions and the quantitative analysis of cDNA microarray images*“ **Journal of Biomedical Optics** (1997), Band 2, Nummer 4, Seiten 364-374

[Chen2003] D. Chen, W.M. Toone, J. Mata, R. Lyne, G. Burns, K. Kivinen, A. Brazma, N. Jones und J. Bähler: „*Global Transcription Responses of Fission Yeast to Environmental Stress*“ **Molecular and Cellular Biology** (2003), Band 14, Seiten 214-229

[Cheung1999] V.G. Cheung, M. Morley, F. Aguilar und A. Massimi: „*Making and reading microarrays*“ **Nature Genetics** (1999), Band 21, Seiten 15-19

[Chudin2001] E. Chudin, R. Walker, A. Kosaka und S.X. Wu: „*Assessment of the relationship between signal intensities and transcript concentration for Affymetrix GeneChip arrays*“ **Genome Biology** (2001), Band 3, Nummer 1, Seiten research0005.1-0005.10

[Clontech2000] Firma Clontech: „*AtlasTM cDNA Expression Arrays User Manual*“ **BD Bioscience Clontech Technical Paper** (2000)

[Clontech2001] Firma Clontech: „*Allgemeine Produktinformation für Atlasarrays AtlasBR.pdf*“ **BD Bioscience Clontech Technical Paper** (2001)

[Clontech2002A] Firma Clontech: „*AtlasTM Glass Microarrays User Manual*“ **BD Bioscience Clontech Technical Paper** (2002)

[Clontech2002B] Firma Clontech: „*AtlasTM Plastic Microarrays User Manual*“ **BD Bioscience Clontech Technical Paper** (2002)

[DeRisi1997] J.L. DeRisi und V.R. Iver: „*Exploring the metabolic and genetic control of gene expression on a genomic scale*“ **Science** (1997), Band 278, Nummer 5338, Seiten 680-686

[Dong2001] F. Dong, H.T. Allawi, T. Anderson und B.P. Neri: „*Secondary structure prediction and structure-specific sequence analysis of single-stranded DNA*“ **Nucleic Acids Research** (2001), Band 29, Nummer 15, Seiten 3248-3257

[Doninger2003] S.W. Doninger, N. Salomonis, K.D. Dahlquist, K. Vranizan, S.C. Lawlor und B.R. Conklin: „*MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data*“ **Genome Biology** (2003), Band 4, Nummer 7, Seiten 1-12

[Dudoit2000] S. Dudoit, Y.W. Yang und M.J. Callow: „*Statistical methods for identifying differntially expressed genes in replicated cDNA microarray experiments*“ **BerkeleyTechReport** (2000), Nummer 578

[Eickhoff1999] B. Eickhoff, B. Korn, M. Schick und A. Poustka: „*Normalization of array hybridization experiments in differntial gene expression analysis*“ **Nucleic Acids Research** (1999), Band 27, Nummer 22, Seiten e33

[Eisen1998] M.B. Eisen, P.T. Spellman und P.O. Brown: „*Cluster analysis and display of genome-wide expression patterns*“ **Proceedings of the National Acadamy of Science** (1998), Band 95, Seiten 14863-14868

[Ermantraut1997] E. Ermantraut: „*Verfahren zur Herstellung von strukturierten*“ **German Patent DE** (1997), Nummer 197 06 570 C1

[Essevaz-Roulet1997] B. Essevaz-Roulet und U. Bockelmann: „*Mechanical separation of the complementary strands of DNA*“ **Proceedings of the National Acadamy of Science** (1997), Band 94, Seiten 11935-11940

[Fink2002] L. Fink, S. Kohlhoff, M.M. Stein, J. Hanze, F. Grimminger und W. Seeger: „*cDNA array hybridization after laser-assisted microdissection from nonneoplastic tissue*“ **American Journal of Pathology** (2002), Band 160, Nummer 1, Seiten 81-90

[Fodor1991] S.P. Fodor, J.L. Read, M.C. Pirrung, L. Stryer, A.T. Lu und D. Solas: „*Light-directed spatially addressable parallel chemical synthesis*“ **Science** (1991), Band 271, Seiten 767-773

[Garber2001] M. Garber, O.G. Troyanskaya, K. Schluens, S. Petersen, C.M. Perou, R.I. Whyte, R.B. Altman, P.O. Brown und D. Botstein: „*Diversity of gene expression in adenocarcinoma of the lung*“ **Proceedings of the National Acadamy of Science** (2001), Band 98, Nummer 24, Seiten 13784-13789

[Ge2000] H. Ge: „*UPA, a universal protein array system for quantitative detection of protein-protein, protein-DNA, protein-RNA, protein-ligand interactions*“ **Nucleic Acids Research** (2000), Band 28, Nummer 2, Seiten e3

[Gelfand1999] C.A. Gelfand, G.E. Plum, S. Mielewczyk, D.P. Remeta und K.J. Breslauer: „*A quantitative method for evaluating the stabilities of nucleic acids*“ **Proceedings of the National Acadamy of Science** (1999), Band 96, Seiten 6113-6118

[Gerland2002] U. Gerland und J.D. Moroz: „*Physical constraints and functional characteristics of transcription factor-DNA interaction*“ **Proceedings of the National Acadamy of Science** (2002), Band 99, Nummer 19, Seiten 12015-12030

[Golub1999] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, J.R. Downing, M.A. Caligiuri und C.D. Bloomfield: „*Molecular classification of cancer: class discovery and class prediction by gene expression monitoring*“ **Science** (1999), Band 286, Seiten 531-537

[Gygi1999] S.P. Gygi, Y. Rochon und B.R. Franza: „*Correlation between Protein and mRNA Abundance in Yeast*“ **Molecular and Cellular Biology** (1999), Band 19, Nummer 3, Seiten 1720-1730

[Hedenfalk2001] I. Hedenfalk, D. Duggan, Y. Chen, M. Radmacher, M. Esteller, O.P. Kallioniemi, B. Wilfond und A. Borg: „*Gene-expression profiles in hereditary breast cancer*“ **New England Journal of Medicine** (2001), Band 344, Nummer 8, Seiten 539-548

[Herwig2001] R. Herwig, P. Aanstad und M. Clark: „*Statistical evaluation of differential expression on cDNA nylon arrays with replicated experiments*“ **Nucleic Acids Research** (2001), Band 29, Nummer 23, Seiten e117

[Hippo2002] Y. Hippo, H. Taniguchi, S. Tsutsumi und N. Machida: „*Global gene expression analysis of gastric cancer by oligonucleotide microarrays*“ **Cancer Research** (2002), Band 62, Seiten 233-240

[Hofacker1994] I.L. Hofacker, W. Fontana, P.F. Stadler und L.S. Bonhoeffer: „*Fast Folding and Comparison of RNA Secondary Structures*“ **Chemical Monthly** (1994), Band 125, Seiten 167-188

[Hoffmann2002] M. Hoffmann, D. Pohlers, D. Koczan, T.C. Kroll, S. Wölfl und R.W. Kinne: „*Gene expression profiles in disease: From gene expression in isolated cells to gene expression in whole tissues and vice versa*“ **European Conference for Computational Biology Saarbrücken** (2002), Nummer P60

[Hoheisel1998] J.D. Hoheisel: „*DNS-Chip-Technologie*“ **Biospektrum** (1998), Band 4, Nummer 6, Seiten 17-20

[Huang2001] Y. Huang, M. Prasad, W.J. Lemon, H. Hampel, F.A. Wright, K. Kornacker, V. LiVolsi, W. Frankel, R.T. Kloos, C. Eng, N.S. Pellegata und A. de la Chapelle: et al. :„*Gene expression in papillary thyroid carcinoma reveals highly consistent profiles*“ **Proceedings of the National Acadamy of Science** (2001), Band 98, Nummer 26, Seiten 15044-9

[IHGSC2001] International Human Genome Sequencing Consortium, E.S. Lander, L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle und W. FitzHugh: et al. :„*Initial sequencing and and analysis of the human genom*“ **Nature** (2001), Band 409, Seiten 860-921

[Jensen1998] K.K. Jensen, H. Orum und P.E. Nielsen: „*Kinetics for hybridization of peptide nucleic acids (PNA) with DNA and RNA studied with the BIAcore technique*“ **Biochemistry** (1997), Band 36, Seiten 5072-5077

[Kafri2000] Y. Kafri und D. Mukamel: „*Why is the DNA denaturation transition first order?*“ **Physical Review Letters** (2000), Band 85, Nummer 23, Seiten 4988-4991

[Kaiser2002] T. Kaiser: „*Ultimate Hardware Tools for Standardization of Microarray Experiments*“ **Cambridge Healthtech Institute - Microarray Dataanalysis - Washington** (2002)

[Karger1993] A.E. Karger, R. Weiss und R.F. Gesteland: „*Line scanning system for direct digital chemiluminescence imaging of DNA sequencing blots.*“ **Analytical Chemistry** (1993), Band 65, Nummer 13, Seiten 1785-93

[Kawakami2001] J. Kawakami, H. Kamiya, K. Yasuda und H. Fujiki: „*Thermodynamic stability of base pairs between 2-hydroxyadenine and incoming nucleotides as a determinant of nucleotide incorporation specificity during replication*“ **Nucleic Acids Research** (2001), Band 29, Nummer 16, Seiten 3289-3296

[Kerr2001A] M.K. Kerr: „*Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments*“ **Proceedings of the National Acadamy of Science** (2001), Band 98, Nummer 16, Seiten 8961-8965

[Kerr2001B] M.K. Kerr und G.A. Churchill: „*Statistical Design and the Analysis of Gene Expression*“ **Genetic Research** (2001), Band 77, Nummer 2, Seiten 123-8

[Kroll2002A] T. Kroll, L. Odyvanova, J.H. Clement, C. Platzer, A. Naumann, N. Marr, K. Höffken und S. Wölfl: „*Molecular characterization of breast cancer cell lines by expression profiling*“ **The Journal of Cancer Research and Clinical Oncology** (2002), Band 128, Seiten 125-134

[Kroll2002B] T. Kroll und S. Wölfl: „*Ranking: a closer look on globalisation methods for normalisation of gene expression arrays*“ **Nucleic Acids Research** (2002), Seiten in press

[Liew1994] C.C. Liew, D.M. Hwang, Y.W. Fung und C. Laurenssen: „*A catalogue of genes in the cardiovascular system as identified by expressed sequence tags*“ **Proceedings of the National Acadamy of Science** (1994), Band 91, Seiten 10645-10649

[Lippincott-Schwartz2001] J. Lippincott-Schwartz, E. Snapp und A. Kenworthy: „*Studying Protein Dynamics In Living Cells*“ **Nature Reviews: Molecular Cell Biology** (2001), Band 2, Seiten rotein

[Lipshutz1999] R.J. Lipshutz, S.P. Fodor und T.R. Gingeras: „*High density synthetic oligonucleotide arrays*“ **Nature Genetics** (1999), Band 21Suppl, Seiten 20-24

[Lockhart1996] D.J. Lockhart, H. Dong, M.C. Byrne, M.T. Follettie, M.V. Gallo, M.S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton und Brown: et al. :„*Expression monitoring by hybridization to high-density oligonucleotide arrays.*“ **Nature Biotechnology** (1996), Band 14, Seiten 1675-1680

[Machl2002] A.W. Machl und C. Schaab: „*Improving DNA array data quality by minimising 'neighbourhood' effects*“ **Nucleic Acids Research** (2002), Band 30, Nummer 22, Seiten e127

[Manduchi2002] E. Manduchi, L.M. Scearce, J.E. Brestelli, G.R. Grant, K.H. Kaestner und C.J. Jr. Stoeckert: „*Comparison of different labeling methods for two-channel high-density microarray experiments*“ **Physiological Genomics** (2002), Band 10, Seiten 169-79

[Masters2000] J.R.W. Masters: „*Human cancer cell lines: fact and fantasy*“ **Nature Reviews: Molecular Cell Biology** (2000), Band 1, Seiten 233-36

[Mathieu-Daude1996] F. Mathieu-Daude, J. Welsh und T. Vogt: „*DNA rehybridization during PCR: the 'Cot effect' and its consequences*“ **Nucleic Acids Research** (1996), Band 24, Nummer 11, Seiten 2080-2086

[Mukamel2002] E.A. Mukamel: „*Phase diagram for unzipping DNA with long-range interactions*“ **Physical Reviews E** (2002), Band 66, Seiten 032901-4

[NatureGenetics1999] „*A note on nomenclature*“ **Nature Genetics** (1999), Band 21Suppl

[Peppel2003] J.vd Peppel, P. Kemmeren, H.v. Bakel, M. Radonjic, D.v. Leenen und F.C.P. Holstege: „*Monitoring global messenger RNA changes in externally controlled microarray experiments*“ **EMBO Reports** (2003), Band 4, Nummer 4, Seiten 387-93

[Perlette2001] J. Perlette: „*Real-time monitoring of intracellular mRNA hybridization inside single living cells*“ **Analytical Chemistry** (2001), Band 73, Seiten 5544-5550

[Perou1999] C.M. Perou, S.S. Jeffrey, M. vandeRijn, C.A. Rees, S.X. Zhu, J.C.F. Lee, D. Lashkari, D. Shalon und P.O. Brown: „*Distinctive gene expression patterns in human mammary epithelial cells and breast cancers*“ **Proceedings of the National Acadamy of Science** (1999), Band 96, Seiten 9212-9217

[Perou2000] C.M. Perou, T. Sorlie, M.B. Eisen, M. vandeRijn, H. Johnsen, L.A. Akslen, E. Fluge, A. Pergamenschikov, C. Williams, S.X. Zhu, P.E. Lonning, A.L. Borresen-Dale und P.O. Brown: et al. :„*Molecular portraits of human breast tumours*“ **Nature** (2000), Band 406, Seiten 747-752

[Peyret1999] N. Peyret, P.A. Seneviratne und H.T. Allawi: „*Nearest-neighbor thermodynamics and NMR of DNA sequences with internal A.A, C.C, G.G, and T.T mismatches*“ **Biochemistry** (1999), Band 38, Seiten 3468-3477

[Phimister1999] B. Phimister: „*Going global (& a note on nomenclature)*“ **Nature Genetics** (1999), Band 21Suppl

[Raytest2002AAM] Firma Raytest: „*Aida Array Metrix*“ **RaytestManual** (2002)

[Rocke2001] D.M. Rocke: „*A model for measurement error for gene expression arrays*“ **The Journal of Computational Biology** (2001), Band 8, Nummer 6, Seiten 557-569

[Sabahi2001] Sabahi A, J. Guidry, G.B. Inamati und M. Manoharan: „*Hybridization of 2'-ribose modified mixed-sequence oligonucleotides: thermodynamic and kinetic studies*“ **Nucleic Acids Research** (2001), Band 29, Nummer 10, Seiten 2163-70

[SantaLucia1996] J. SantaLucia und H.T. Allawi: „*Improved nearest-neighbor parameters for predicting DNA duplex stability*“ **Biochemistry** (1996), Band 35, Seiten 3555-3562

[SantaLucia1998] J. SantaLucia: „*A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics*“ **Proceedings of the National Acadamy of Science** (1998), Band 95, Seiten 1460-1465

[Schena1995] M. Schena, D. Shalon, R.W. Davis und P.O. Brown: „*Quantitative monitoring of gene expression patterns with a complementary DNA microarray.*“ **Science** (1995), Band 270, Seiten 467-470

[Schena1996] M. Schena, D. Shalon, R. Heller, A. Chai, P.O. Brown und R.W. Davis: „*Parallel human genome analysis: microarray-based expression monitoring of 1000 genes*“ **Proceedings of the National Acadamy of Science** (1996), Band 93, Nummer 20, Seiten 10614-10619

[Schuchardt2000] J. Schuchardt, D. Beule, A. Malik und E. Wolski: „*Normalization strategies for cDNA microarrays*“ **Nucleic Acids Research** (2000), Band 28, Nummer 10, Seiten e47

[Schwille1996] P. Schwille und F. Oehlenschlager: „*Quantitative hybridization kinetics of DNA probes to RNA in solution followed by diffusional fluorescence correlation analysis*“ **Biochemistry** (1996), Band 35, Seiten 10182-10193

[Sorlie2001] T. Sorlie, C.M. Perou, R. Tibshirani, T. Aas, M. van de Rijn, S.S. Jeffrey, T. Thorsen, H. Quist, J.C. Matese, P.O. Brown, D. Botstein und P. Eystein Lonning: et al. :„*Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications*“ **Proceedings of the National Acadamy of Science** (2001), Band 98, Nummer 19, Seiten 10869-10874

[Southern1974] E.M. Southern: „*Detection of Specific Sequences Among DNA-Fragments Separeted by Gel Electrophoresis*“ **Journal of Molecular Biology** (1974), Band 98, Seiten 505-517

[Southern1999] E. Southern und K. Mir: „*Molecular interactions on microarrays*“ **Nature Genetics** (1999), Band 21, Seiten 5-9

[Spanakis1993] E. Spanakis: „*Problems related to the interpretation of autoradiographic data on gene expression using common constitutive transcripts as controls*“ **Nucleic Acids Research** (1993), Band 21, Nummer 16, Seiten 3809-3819

[Stein2001] L. Stein: „*Genome Annotation: From Sequence To Biology*“ **Nature Reviews: Genetics** (2001), Band 2, Seiten 493-505

[Sugimoto1996] N. Sugimoto, S. Nakano und M. Yoneyama: „*Improved thermodynamic parameters and helix initiation factor to predict stability of DNA duplexes*“ **Nucleic Acids Research** (1996), Band 24, Nummer 22, Seiten 4501-4505

[Tukey1962] J.W. Tukey: „“ **Annals of Mathematical Statistics** (1962), Band 33, Seiten 1-67

[Vandesompele2002] J. Vandesompele, K. De Preter, F. Pattyn und B. Poppe: „*Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes.*“ **Genome Biology** (2002), Band 3, Nummer 7, Seiten 0034.1-12

[Velculescu1995] V.E. Velculescu, L. Zhang, B.B. Vogelstein und K.W. Kinzler: „*Serial analysis of gene expression.*“ **Science** (1995), Band 270, Seiten 484-487

[Venter2001] C. Venter: „*The sequence of the human genome*“ **Science** (2001), Band 291, Seiten 1304-1351

[Vernon2000] S.D. Vernon, E.R. Unger, M. Rajeevan und I.M. Dimulescu: „*Reproducibility of Alternative Probe Synthesis Approaches for Gene Expression Profiling with Arrays*“ **Journal of Molecular Diagnostics** (2000), Band 2, Nummer 3, Seiten 124-127

[Voelker2001] J. Völker und H.H. Klump: „*Communication between noncontacting macromolecules*“ **Proceedings of the National Acadamy of Science** (2001), Band 98, Seiten 7694-7699

[Welford1998] S.M. Welford, J. Gregg, E. Chen und D. Garrison: „*Detection of differentially expressed genes in primary tumor tissues using representational differences analysis coupled to microarray hybridization*“ **Nucleic Acids Research** (1998), Band 26, Nummer 12, Seiten 3059-3065

[Welsh2001] J.B. Welsh, P.P. Zarrinkar, L.M. Sapinoso und S.G. Kern: „*Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer*“ **Proceedings of the National Acadamy of Science** (2001), Band 98, Nummer 3, Seiten 1176-1181

[Wodicka1997] L. Wodicka, H. Dong, M. Mittmann und M. Ho: „*Genome-wide expression monitoring in Saccharomyces cerevisiae*“ **Nature Biotechnology** (1997), Band 15, Seiten 1359-1367

[Woelfl2000] S. Wölfl: „*BioChip-Technologien*“ **transkript Laborwelt** (2000), Nummer 3, Seiten 12-20

[Workman2002] C. Workman, L. Jensen, H. Jarmer, R. Berka, L. Gautier, H. Nielser, H.H. Saxild, C. Nielsen, S. Brunak und S. Knudsen: „*A new non-linear normalization method for reducing variability in DNA microarray experiments*“ **Genome Biology** (2002), Band 3, Seiten research0048.1-16

[Xia1998] T. Xia, J. SantaLucia, M.E. Burkard und R. Kierzek: „*Thermodynamic Parameters for an Expanded Nearest-Neighbor Model for Formation of RNA Duplexes with Watson-Crick Base Pairs*“ **Biochemistry** (1998), Band 37, Seiten 14719-14735

[Yamaguchi2001] S. Yamaguchi, M. Kobayashi, S. Mitsui und Y. Ishida: „*View of a mouse clock gene ticking*“ **Nature** (2001), Band 409, Seiten 684

[Yang2000] Y.H. Yang, M.J. Buckley und S. Dudoit: „*Comparison of methods for image analysis on cDNA microarray data*“ **BerkeleyTechReport** (2000), Nummer 584

[Yang2002A] Y.H. Yang, S. Dudoit, P. Luu und D.M. Lin: „*Normalization for cDNA microarray data: a robust composite method addresing single and multiple slide systemic variation*“ **Nucleic Acids Research** (2002), Band 30, Nummer 4, Seiten e15

[Yang2002B] Y.H. Yang und T. Speed: „*Design Issues for Microarray Experiments*“ **Nature Reviews: Genetics** (2002), Band 3, Seiten 579-588

[YangMC2001] M.C.K. Yang, Q.G. Ruan, J.J. Yang und S. Eckenrode: „*A statistical method for flagging weak spots improves normalization and ratio estimates in microarrays*“ **Physiological Genomics** (2001), Band 7, Seiten 45-53

[YangYH2001] Y.H. Yang, S. Dudoit und P. Luu: „*Normalization for cDNA Microarray Data*“ **BerkeleyTechReport** (2001), Nummer 589

[Zien2001] A. Zien, T. Aigner und R. Zimmer: „*Centralization: a new method for the normalization of gene expression data*“ **Bioinformatics** (2001), Band 17, Seiten s323-s331

[2] Verwendete Lehrbücher

[Ackermann] T. Ackermann: „*Physikalische Biochemie*“, 1.Auflage, **Springer** Berlin... (1992)

[Adam] G. Adam, P. Läuger, G. Stark: „*Physikalische Chemie und Biophysik*“, 3.Auflage, **Springer** Berlin... (1995)

[Alberts] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, J.D. Watson: „*Molecular Biology of the Cell*“, 3.Auflage, **Garland** New York & London (1996)

[Bronstein] I.N. Bronstein, K.A. Semendjajew, G. Musiol, H. Mühlig: „*Taschenbuch der Mathematik*“ 4.Auflage, **Verlag Harri Deutsch Thun** Frankfurt/M. (1999)

[Cantor] C.R. Cantor, P.R. Schimmel: „*Biophysical Chemistry - Part III - The behavior of biological macromolecules*“, 10. unveränderte Auflage,

[Dawkins1996] R. Dawkins „*Climbing Mount Improbable*“, 1.Auflage, **Viking** London (1996) **Freeman** New York 1980 (1998)

[Hayat] M.A. Hayat (Ed.): „*Immunogold-Silver Staining, Principles, Methods, and Applications*“, CRC Press (1995)

[KoehlerBS] W. Köhler, G. Schachtel, P. Voleske: „*Biostatistik*“ 3.Auflage, **Springer** Berlin... (2002)

[KoehlerPC] P.W. Köhler:„*Physikalische Chemie*“ 1.Auflage (2.Nachdruck) , **VHC** Weinheim 1990

[Lehninger] A.L. Lehninger, D.L. Nelson, M.M. Cox: „*Prinzipien der Biochemie*“, 2.Auflage, **Spektrum** Heidelberg Berlin Oxford (1998)

[Lewin] B. Lewin: „*Molekularbiologie der Gene*“, 6.Auflage, Spektrum Heidelberg Berlin Oxford (1998)

[Lottspeich] F. Lottspeich, H. Zorbas: „*Bioanalytik*“, 1.Auflage, **Spektrum** Heidelberg Berlin Oxford (1998)

[Sachs] Sachs L „*Angewandte Statistik*“, 9.Auflage, **Springer** Berlin... (1999)

[StatSoft] StatSoft, Inc. „*Electronic Statistics Textbook.*“, Tulsa (2002), OK: StatSoft. WEB:

[StoeckerMat] H. Stöcker (Hrsg.): „*Taschenbuch mathematischer Formeln und moderner Verfahren*“ 2.Auflage, **Verlag Harri Deutsch** Thun Frankfurt/M. (1993)

[StoeckerPh] H. Stöcker (Hrsg.): „*Taschenbuch der Physik*“ 3.Auflage, **Verlag Harri Deutsch** Thun Frankfurt/M. (1998)

[Trampisch] H.J. Trampisch, J. Windeler:„*Medizinische Statistik*“ 2.Auflage, **Springer** Berlin... (2000)

[VoetVoet] D. Voet, J.G. Voet:„*Biochemistry*“, 1.Auflage, **Wiley** NewYork (1995)
`http://www.statsoft.com/textbook/stathome.html`

[3] Linkangaben zu zitierten Webseiten oder weiterführenden Informationen

[ATCC] **American Type Culture Collection** www.atcc.org

[Affymetrix] **Affymetrix Inc.** www.affymetrix.com

[Clondiag] **Clondiag Chip Technologies Jena GmbH** www.clondiag.com

[Clontech] **BD Bioscience Inc.**:Clontech www.clontech.com

[GB:NM001200] **Genbankeintrag** http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=nucleotide&list_uids=4557368&dopt=GenBank

[Mathworks] **Mathworks Inc.** www.mathworks.com

[MGED] **Microarray Gene Expression Data Group** www.mged.org

[Microsoft] **Microsoft GmbH** www.microsoft.com

[PubMed] Literaturdatenbank des US-amerikanischen **National Institutes of Health** http://www.ncbi.nlm.nih.gov/PubMed/

[Raytest] **Raytest GmbH** www.raytest.de

Anhang A

8.1 Beispielsequenzen für die Hybridisierungsmodelle

8.1.1 BMP2-Sequenz [GB:NM001200]

$BMP2 - mRNA$:

5'GGG GAC TTC TTG AAC TTG CAG GGA GAA TAA CTT GCG CAC CCC ACT TTG CGC CGG TGC CTT TGC CCC
AGC GGA GCC TGC TTC GCC ATC TCC GAG CCC CAC CGC CCC TCC ACT CCT CGG CCT TGC CCG ACA CTG AGA
CGC TGT TCC CAG CGT GAA AAG AGA GAC TGC GCG GCC GGC ACC CGG GAG AAG GAG GAG GCA AAG AAA AGG
AAC GGA CAT TCG GTC CTT GCG CCA GGT CCT TTG ACC AGA GTT TTT CCA TGT GGA CGC TCT TTC AAT GGA
CGT GTC CCC GCG TGC TTC TTA GAC GGA CTG CGG TCT CCT AAA GGT CGA CCA TGG TGG CCG GGA CCC GCT
GTC TTC TAG CGT TGC TGC TTC CCC AGG TCC TCC TGG GCG GCG CGG CTG GCC TCG TTC CGG AGC TGG GCC
GCA GGA AGT TCG CGG CGG CGT CGT CGG GCC GCC CCT CAT CCC AGC CCT CTG ACG AGG TCC TGA GCG AGT
TCG AGT TGC GGC TGC TCA GCA TGT TCG GCC TGA AAC AGA GAC CCA CCC CCA GCA GGG ACG CCG TGG TGC
CCC CCT ACA TGC TAG ACC TGT ATC GCA GGC ACT CAG GTC AGC CGG GCT CAC CCG CCC CAG ACC ACC GGT
TGG AGA GGG CAG CCA GCC GAG CCA ACA CTG TGC GCA GCT TCC ACC ATG AAG AAT CTT TGG AAG AAC TAC
CAG AAA CGA GTG GGA AAA CAA CCC GGA GAT TCT TCT TTA ATT TAA GTT CTA TCC CCA CGG AGG AGT TTA
TCA CCT CAG CAG AGC TTC AGG TTT TCC GAG AAC AGA TGC AAG ATG CTT TAG GAA ACA ATA GCA GTT TCC
ATC ACC GAA TTA ATA TTT ATG AAA TCA TAA AAC CTG CAA CAG CCA ACT CGA AAT TCC CCG TGA CCA GAC
TTT TGG ACA CCA GGT TGG TGA ATC AGA ATG CAA GCA GGT GGG AAA GTT TTG ATG TCA CCC CCG CTG TGA
TGC GGT GGA CTG CAC AGG GAC ACG CCA ACC ATG GAT TCG TGG TGG AAG TGG CCC ACT TGG AGG AGA AAC
AAG GTG TCT CCA AGA GAC ATG TTA GGA TAA GCA GGT CTT TGC ACC AAG ATG AAC ACA GCT GGT CAC AGA
TAA GGC CAT TGC TAG TAA CTT TTG GCC ATG ATG GAA AAG GGC ATC CTC TCC ACA AAA GAG AAA AAC GTC
AAG CCA AAC ACA AAC AGC GGA AAC GCC TTA AGT CCA GCT GTA AGA GAC ACC CTT TGT ACG TGG ACT TCA
GTG ACG TGG GGT GGA ATG ACT GGA TTG TGG CTC CCC CGG GGT ATC ACG CCT TTT ACT GCC ACG GAG AAT
GCC CTT TTC CTC TGG CTG ATC ATC TGA ACT CCA CTA ATC ATG CCA TTG TTC AGA CGT TGG TCA ACT CTG
TTA ACT CTA AGA TTC CTA AGG CAT GCT GTG TCC CGA CAG AAC TCA GTG CTA TCT CGA TGC TGT ACC TTG
ACG AGA ATG AAA AGG TTG TAT TAA AGA ACT ATC AGG ACA TGG TTG TGG AGG GTT GTG GGT GTC GCT AGT
ACA GCA AAA TTA AAT ACA TAA ATA TAT ATA TA 3'

$BMP2 - cDNA$:

5'TAT ATA TAT ATT...(reverse komplementär zur mRNA)

8.1.2 BMP2-Sonden unterschiedlicher Länge

$B500$:

5'CTA AAG GTC GAC CAT GGT GGC CGG GAC CCG CTG TCT TCT AGC GTT GCT GCT TCC CCA GGT CCT CCT
GGG CGG CGC GGC TGG CCT CGT TCC GGA GCT GGG CCG CAG GAA GTT CGC GGC GGC GTC GTC GGG CCG CCC
CTC ATC CCA GCC CTC TGA CGA GGT CCT GAG CGA GTT CGA GTT GCG GCT GCT CAG CAT GTT CGG CCT GAA
ACA GAG ACC CAC CCC CAG CAG GGA CGC CGT GGT GCC CCC CTA CAT GCT AGA CCT GTA TCG CAG GCA CTC
AGG TCA GCC GGG CTC ACC CGC CCC AGA CCA CCG GTT GGA GAG GGC AGC CAG CCG AGC CAA CAC TGT GCG

CAG CTT CCA CCA TGA AGA ATC TTT GGA AGA ACT ACC AGA AAC GAG TGG GAA AAC AAC CCG GAG ATT CTT
CTT TAA TTT AAG TTC TAT CCC CAC GGA GGA GTT TAT CAC CTC AGC AGA GCT TCA GGT TTT CCG AGA ACA
GAT GCA AGA TGC TTT AGG AA 3'

$B200$:
5'CCC CCA GCA GGG ACG CCG TGG TGC CCC CCT ACA TGC TAG ACC TGT ATC GCA GGC ACT CAG GTC AGC
CGG GCT CAC CCG CCC CAG ACC ACC GGT TGG AGA GGG CAG CCA GCC GAG CCA ACA CTG TGC GCA GCT TCC
ACC ATG AAG AAT CTT TGG AAG AAC TAC CAG AAA CGA GTG GGA CAA CCC GGA GAT TCT TCT TT 3'

$B100$:
5'GGG CTC ACC CGC CCC AGA CCA CCG GTT GGA GAG GGC AGC CAG CCG AGC CAA CAC TGT GCG CAG CTT
CCA CCA TGA AGA ATC TTT GGA AAA AGA ACT ACC AGA A 3'

$B50$:
5'GAG GGC AGC CAG CCG AGC CAA CAC TGT GCG CAG CTT CCA CCA TGA AGA AT 3'

$B20$:
5'GCG CAG CTT CCA CCA TGA AG 3'

$B10$:
5'CGG AGG AGT T 3'

$B25_{pmCG}$:
5'AAG TTC TAT CCC <C:G>AC GGA GGA GTT T 3'

$B25_{mmCA}$:
5'AAG TTC TAT CCC <A:G>AC GGA GGA GTT T 3'

$B25_{mmCC}$:
5'AAG TTC TAT CCC <G:G>AC GGA GGA GTT T 3'

$B25_{mmCT}$:
5'AAG TTC TAT CCC <T:G>AC GGA GGA GTT T 3'

8.1.3 willkürliche Sonden mit unterschiedlichem GC-Gehalt

$S30_{GC100\%}$:
5'GCG GGC CGC CCG CCG CGC CCC GCC GGG GCG 3'

$S30_{GC80\%}$:
5'GAG GGC GGC CAG CCG CGC CAG CAC CGT GCA 3'

$S30_{GC60\%}$:
5'GAG GGC AGC CAG CTG AGC CAA CAC TGT GAT 3'

$S30_{GC40\%}$:
5'GAG AAC ATC CAA CTT AGC AAA CAC TTT GCA 3'

$S30_{GC20\%}$:
5'GAT AGA ATC TAA TAT AGT AAA CAA TTT TCA 3'

$S30_{GC0\%}$:
`5'TAT TTA ATA AAT AAT ATA AAT AAA TTT TAT 3'`

8.2 Ableitung des 1.Modells

Einfaches Modell der Hybridisierung einer markierten cDNA/mRNA-Spezies A mit der korrespondierenden Sonde S ohne Annahme von Kreuzhybridisierungen. A kann sich im gesamten Volumen der Probe V_{probe} bewegen. S und das Reaktionsprodukt AS kann sich aufgrund der Immobilisierung der Sonde nur in einem begrenzten Volumen V_{sonde} bewegen. Reaktionszeit ist lang genug, daß sich das Gleichgewicht der Reaktion einstellt.

Definitionen
V_{total} . . .Gesamtvolumen

$$V_{total} = V_{sonde} + V_{probe} \tag{8.1}$$

n_{AS}, n_S, n_A . . .Stoffmengen der beteiligten Stoffe
n_{A0}, n_{S0} . . .Gesamtstoffmenge der unreagierten Ausgangsstoffe A und S

$$n_{A0} = n_A + n_{AS} \tag{8.2}$$
$$n_{S0} = n_S + n_{AS} \tag{8.3}$$

c_{AS}, c_S, c_A . . .Konzentrationen der beteiligten Stoffe

$$c_A = \frac{n_A}{V_{total}} \tag{8.4}$$
$$c_S = \frac{n_S}{V_{total}} \tag{8.5}$$
$$c_{AS} = \frac{n_{AS}}{V_{sonde}} \tag{8.6}$$

K . . .Gleichgewichtskonstante der Reaktion

$$K = \frac{c_{AS}}{c_A c_S} = \frac{n_{AS}}{n_A n_S} \cdot \frac{V_{sonde} \cdot V_{total}}{V_{sonde}} \tag{8.7}$$

f . . .Hilfsvariable

$$f = \frac{n_A n_S}{n_{AS}} = \frac{V_{total}}{K} \tag{8.8}$$

Ableitungen
aus den obigen Gleichungen folgt:

$$n_{A0} = f\frac{n_{AS}}{n_S} + n_{AS} \tag{8.9}$$
$$n_{A0} = \left(\frac{f}{(n_{S0} - n_{AS})} + 1\right) n_{AS} \tag{8.10}$$

Umstellen der Gleichung nach n_{AS}, der detektierbaren Größe ergibt:

$$n_{AS} = \frac{1}{2}n_{A0} + \frac{1}{2}\frac{V_{total}}{K} + \frac{1}{2}n_{S0} - \frac{1}{2}\sqrt{n_{A0}^2 + 2n_{A0}\frac{V_{total}}{K} - 2n_{A0}n_{S0} + \left(\frac{V_{total}}{K}\right)^2 + 2n_{S0}\frac{V_{total}}{K} + n_{S0}^2} \tag{8.11}$$

8.3 Ableitung des 2.Modells - Kompetetive Hybridisierung

Einfaches Modell der Hybridiserung zweier markierter cDNA/mRNA-Spezies A und B mit der korrespondierenden Sonde S ohne Annahme von Kreuzhybridisierungen (keine Bildung von AB). A und B können sich im gesamten Volumen der Probe V_{probe} bewegen. S und die Reaktionsprodukte AS und BS können sich aufgrund der Immobilisierung der Sonde nur in einem begrenzten Volumen V_{sonde} bewegen. Reaktionszeit ist lang genug, daß sich das Gleichgewicht der Reaktion einstellt.

Definitionen

V_{total} ...Gesamtvolumen

$$V_{total} = V_{sonde} + V_{probe} \tag{8.12}$$

$n_{AS}, n_A, n_{BS}, n_B, n_S$ ...Stoffmengen der beteiligten Stoffe
n_{A0}, n_{B0}, n_{S0} ...Gesamtstoffmenge der unreagierten Ausgangsstoffe A, B und S

$$n_{A0} = n_A + n_{AS} \tag{8.13}$$
$$n_{B0} = n_B + n_{BS} \tag{8.14}$$
$$n_{S0} = n_S + n_{AS} + n_{BS} \tag{8.15}$$

$c_{AS}, c_A, c_{BS}, c_B, c_S$ ...Konzentrationen der beteiligten Stoffe

$$c_A = \frac{n_A}{V_{total}} \tag{8.16}$$
$$c_B = \frac{n_B}{V_{total}} \tag{8.17}$$
$$c_S = \frac{n_S}{V_{sonde}} \tag{8.18}$$
$$c_{AS} = \frac{n_{AS}}{V_{sonde}} \tag{8.19}$$
$$c_{BS} = \frac{n_{BS}}{V_{sonde}} \tag{8.20}$$

K_A, K_B ...Gleichgewichtskonstanten der beiden Teilreaktion

$$K_A = \frac{c_{AS}}{c_A c_S} = \frac{n_{AS}}{n_A n_S} \cdot \frac{V_{sonde} \cdot V_{total}}{V_{sonde}} \tag{8.21}$$
$$K_B = \frac{c_{BS}}{c_B c_S} = \frac{n_{BS}}{n_B n_S} \cdot \frac{V_{sonde} \cdot V_{total}}{V_{sonde}} \tag{8.22}$$

f_a, f_b ...Hilfsvariablen

$$f_a = \frac{V_{total}}{K_A} = \frac{n_A n_S}{n_{AS}} \tag{8.23}$$
$$f_b = \frac{V_{total}}{K_B} = \frac{n_B n_S}{n_{BS}} \tag{8.24}$$

Ableitungen

aus den obigen Gleichungen folgt:

$$n_{A0} = \left(\frac{f_a}{(n_{S0} - n_{AS} - n_{BS})} + 1 \right) n_{AS} = \left(\frac{f_a}{n_S} + 1 \right) n_{AS} \tag{8.25}$$
$$n_{AS} = \frac{n_{A0} n_S}{f_a + n_S} \tag{8.26}$$
$$n_{B0} = \left(\frac{f_a}{(n_{S0} - n_{AS} - n_{BS})} + 1 \right) n_{BS} = \left(\frac{f_b}{n_S} + 1 \right) n_{BS} \tag{8.27}$$
$$n_{BS} = \frac{n_{B0} n_S}{f_b + n_S} \tag{8.28}$$

diese Gleichungen ergeben:

$$n_{S0} = n_S + \frac{n_{A0} n_S}{f_a + n_S} + \frac{n_{B0} n_S}{f_b + n_S} \tag{8.29}$$

Diese Gleichung wiederum läßt sich in die Normalform einer Gleichung 3. Ordnung umstellen:

$$\begin{aligned} 0 &= n_S^3 + (f_a + f_b + n_{A0} + n_{B0} - n_{S0})\, n_S^2 + \\ &\quad (f_a f_b + f_b n_{A0} + f_a n_{B0} - f_b n_{S0} - f_a n_{S0})\, n_S - f_a f_b n_{S0} \end{aligned} \tag{8.30}$$

$$0 = n_S^3 + p_1 n_S^2 + p_2 n_S + p_3 \tag{8.31}$$

Die vollständige Umstellung ist zwar möglich aber nicht notwendig, da ich für die Berechnung von n_S den polynomischen Lösung von MatLab benutzte. Die benötigten Koeffizienten für Gleichung 8.31 lauten:

$$p_1 = f_a + f_b + n_{A0} + n_{B0} - n_{S0} \tag{8.32}$$

$$p_2 = f_a f_b + f_b n_{A0} + f_a n_{B0} - f_b n_{S0} - f_a n_{S0} \tag{8.33}$$

$$p_3 = -f_a f_b n_{S0} \tag{8.34}$$

Die Auswahl der drei möglichen Lösungen erfolgt dadurch, daß nur die Lösung verwendet wird, die folgende Eigenschaften hat:

$$n_S \geq 0 \tag{8.35}$$

$$n_S \leq n_{S0} \tag{8.36}$$

$$n_S \in \Re \tag{8.37}$$

Letztere Bedingung ist leider durch Rechenungenauigkeiten nicht immer erfüllt, d.h. es gibt auch Fälle in denen alle Lösungen von n_S komplexe Anteile haben. In diesen Fall wurde der Realteil der Lösung genommen, die die ersten zwei Bedingungen erfüllt und den kleinsten Komplexteil hat. Die eigentlich interessanten Größen sind n_{AS} und n_{BS}, da sie meßbar sind. Sie können mittels bekannten n_S durch die Gleichungen 8.26 und 8.28 berechnet werden.

8.4 Ableitung des 3.Modells - Hybridisierung an zwei Sonden

Einfaches Modell der Hybridiserung einer markierten cDNA/mRNA-Spezies A mit der korrespondierenden Sonde S_{pm}[1] und einer alternativen Sonde S_{mm}[2]. A kann sich im gesamten Volumen der Probe V_{probe} bewegen. S und das Reaktionsprodukt AS kann sich aufgrund der Immobilisierung der Sonde nur in einem begrenzten Volumen V_{sonde} bewegen. Reaktionszeit ist lang genug, daß sich das Gleichgewicht der Reaktion einstellt.

Definitionen

V_{total} . . .Gesamtvolumen

$$V_{total} = V_{sonde} + V_{probe} \tag{8.38}$$

$n_{AS_{pm}}, n_{S_{pm}}, n_{AS_{mm}}, n_{S_{mm}}, n_A$ . . .Stoffmengen der beteiligten Stoffe
$n_{S_{pm}0}, n_{S_{mm}0}, n_{A0}$ . . .Gesamtstoffmenge der unreagierten Ausgangsstoffe A, B und S

$$n_{S_{pm}0} = n_{S_{pm}} + n_{AS_{pm}} \tag{8.39}$$

$$n_{S_{mm}0} = n_{S_{mm}} + n_{AS_{mm}} \tag{8.40}$$

$$n_{A0} = n_A + n_{AS_{pm}} + n_{AS_{mm}} \tag{8.41}$$

[1] pm . . . perfect match
[2] mm . . . mismatch

$c_{AS_{pm}}, c_{S_{pm}}, c_{AS_{mm}}, c_{S_{mm}}, c_A \ldots$Konzentrationen der beteiligten Stoffe

$$c_{S_{pm}} = \frac{n_{S_{pm}}}{V_{sonde}} \tag{8.42}$$

$$c_{S_{mm}} = \frac{n_{S_{mm}}}{V_{sonde}} \tag{8.43}$$

$$c_A = \frac{n_A}{V_{total}} \tag{8.44}$$

$$c_{AS_{pm}} = \frac{n_{AS_{pm}}}{V_{sonde}} \tag{8.45}$$

$$c_{AS_{mm}} = \frac{n_{AS_{mm}}}{V_{sonde}} \tag{8.46}$$

$K_{pm}, K_{mm} \ldots$Gleichgewichtskonstanten der beiden Teilreaktion

$$K_{pm} = \frac{c_{AS_{pm}}}{c_{S_{pm}} c_A} = \frac{n_{AS_{pm}}}{n_{S_{pm}} n_A} \cdot \frac{V_{sonde} \cdot V_{total}}{V_{sonde}} \tag{8.47}$$

$$K_{mm} = \frac{c_{AS_{mm}}}{c_{S_{mm}} c_A} = \frac{n_{AS_{mm}}}{n_{S_{mm}} n_A} \cdot \frac{V_{sonde} \cdot V_{total}}{V_{sonde}} \tag{8.48}$$

$f_a, f_b \ldots$Hilfsvariablen

$$f_a = \frac{V_{total}}{K_{pm}} = \frac{n_{S_{pm}} n_A}{n_{AS_{pm}}} \tag{8.49}$$

$$f_b = \frac{V_{total}}{K_{mm}} = \frac{n_{S_{mm}} n_A}{n_{AS_{mm}}} \tag{8.50}$$

Ableitungen

aus den obigen Gleichungen folgt:

$$n_{S_{pm}0} = \left(\frac{f_a}{\left(n_{A0} - n_{AS_{pm}} - n_{AS_{mm}}\right)} + 1 \right) n_{AS_{pm}} = \left(\frac{f_a}{n_A} + 1 \right) n_{AS_{pm}} \tag{8.51}$$

$$n_{AS_{pm}} = \frac{n_{S_{pm}0} n_A}{f_a + n_A} \tag{8.52}$$

$$n_{S_{mm}0} = \left(\frac{f_a}{\left(n_{A0} - n_{AS_{pm}} - n_{AS_{mm}}\right)} + 1 \right) n_{AS_{mm}} = \left(\frac{f_b}{n_A} + 1 \right) n_{AS_{mm}} \tag{8.53}$$

$$n_{AS_{mm}} = \frac{n_{S_{mm}0} n_A}{f_b + n_A} \tag{8.54}$$

diese Gleichungen ergeben:

$$n_{A0} = n_A + \frac{n_{S_{pm}0} n_A}{f_a + n_A} + \frac{n_{S_{mm}0} n_A}{f_b + n_A} \tag{8.55}$$

Diese Gleichung wiederum läßt sich in die Normalform einer Gleichung 3. Ordnung umstellen:

$$\begin{aligned} 0 = {} & n_A^3 + \left(f_a + f_b + n_{S_{pm}0} + n_{S_{mm}0} - n_{A0}\right) n_A^2 + \\ & \left(f_a f_b + f_b n_{S_{pm}0} + f_a n_{S_{mm}0} - f_b n_{A0} - f_a n_{A0}\right) n_A - f_a f_b n_{A0} \end{aligned} \tag{8.56}$$

Die vollständige Umstellung ist zwar möglich aber nicht notwendig, da ich für die Berechnung von n_A den polynomischen Lösung von MatLab benutzte. Weiterhin ist diese Lösung (Gleichung 8.57) equivalent zum vorherigem Abschnitt (Gleichung 8.31), so daß mit den folgenden Variablentransformationen die gleiche Programmlösung verwendet werden kann.

Eingangsvariablen

$$n_{S_{pm}0} \mapsto n_{A0} \tag{8.57}$$
$$n_{S_{mm}0} \mapsto n_{B0} \tag{8.58}$$
$$n_{A0} \mapsto n_{S0} \tag{8.59}$$
$$V_{total} \mapsto V_{total} \tag{8.60}$$
$$K_{pm} \mapsto K_A \tag{8.61}$$
$$K_{mm} \mapsto K_B \tag{8.62}$$

Die Ausgangsvariablen „mappen“ zurück auf die gewünschten Variablen $n_{AS_{pm}}$,$n_{AS_{pm}}$ und n_A.

Ausgangsvariablen

$$n_{AS} \mapsto n_{AS_{pm}} \tag{8.63}$$
$$n_{BS} \mapsto n_{AS_{mm}} \tag{8.64}$$
$$n_S \mapsto n_A \tag{8.65}$$

8.5 AGM-Funktion

```
function trim=ATM(X,lower,upper);
%ATM The asymetrically trimmed mean of X is a generalized form of estimates of
%   a distribution dependent scaling factor.
%   trim = trimmedmean(X,lower,upper) calculates the mean of X excluding the highest
%   and lowest parts of the data. For matrices, ATM(X) is a vector
%   containing the asymetrically trimmed mean for each column. The scalars, LOWER and UPPER
%   must take values between 0 and 1 and the sum of both must be
%   smaller 1.
%
%   For matrix, X, TRIM = ATM(X,lower,upper); is a row vector containing the
%   asymetrically trimmed mean for each column of X.
%
%   Copyright 2001 Torsten C. Kroll
%   $Date: 2003/07/15 17:32:00 $
if nargin < 3
    error('Three input arguments are required.');
end
P = 1; % change to P=100; if you need limits in percent
if lower+upper > P  | lower < 0 | upper < 0
    if P==1;error('The sum of the lower and upper limits must take values between 0 and 1.');
    else error('The sum of the lower and upper limits must take values between 0 and 100%.'); end;
end
[row col] = size(X);
if row < 2
    error('X must contain at least 2 rows.');
end
X = sort(X);
I = [X(1,:) ; X];%
D = [X; X(end,:)] - I;
%define lower limit
l_rank = lower./P.*row;
l_int  = floor(l_rank + 0.5);
l_frac = (l_rank + 0.5) - l_int;
%define upper limit
u_rank = (P - upper)./P.*row;
u_int  = floor(u_rank + 0.5);
u_frac = (u_rank + 0.5) - u_int;
%define integration distance
dist   = u_rank - l_rank;
```

```
%integrate from 0 to lower part
l_Prctl = I(l_int+1,:)+l_frac*D(l_int+1,:); %lower percentile/quantile
l_Sum   = sum(I(1:(l_int+1),:),1) - ...%;
          0.5.*I(l_int+1,:) + ...%;%
          (l_frac).*I(l_int+1,:) + ...%
          0.5.*l_frac.*l_frac.*D(l_int+1,:);%
%integrate from 0 to upper part
u_Prctl = I(u_int+1,:)+u_frac*D(u_int+1,:); %upper percentile/quantile
u_Sum   = sum(I(1:(u_int+1),:),1) + ...%;
          (u_frac-0.5).*I(u_int+1,:) + ...%;%
          0.5.*u_frac.*u_frac.*D(u_int+1,:);%
% make average by division of integral value by integration distance
if dist<(row/10000); %dist==0 would result in /0 warning but trim is in this case equal to both percentiles
    trim=l_Prctl;
    warning('Too much data trimmed. Result approximated (percentile/quantile).');
else
    trim=(u_Sum-l_Sum)./dist;%
end;
```

Anhang B

Danksagung

Für die langjährige sehr gute Zusammenarbeit möchte ich mich besonders bei meinem Betreuer Prof. Dr. Stefan Wölfl bedanken. Durch seine Unterstützung und Ermutigung wurden aus meinen Ideen diese Arbeit. Aus den Diskussionen gemeinsamer Problemstellungen konnte ich viele Anregungen mitnehmen.

Für weitere interessante Anregungen und Diskussionen möchte ich PD Dr. Reinhard Guthke und Mitgliedern seiner Arbeitsgruppe danken. Dort besonders Dipl.Ing. Daniel Hahn für die Zusammenarbeit bei der Hintergrundbestimmung und Dr. Martin Hoffmann für Diskussionen von Auswerteproblemen.

In unserer Arbeitsgruppe an der Klinik für Innere Medizin möchte ich bei allen Kolleginnen und Kollegen danken. Für die Überlassung experimenteller Daten und diverser „Probleme“ danke ich besonders Dr. Larissa Pusch, Stefan Knoth und Ana Kitanovič. Viele der gemeinsamen Problemstellungen wurden in dieser Arbeit aufgegriffen. Anregende Diskussionen und viele Korrekturhinweise verdanke ich Sotirios Ziagos.

Meinem früheren Chef Prof.Dr. Hans-Peter Saluz und seiner Abteilung für Zell- und Molekularbiologie am Hans-Knöll-Institut danke ich für die Unterstützung bei der der Bearbeitung meines ersten Themas und der Möglichkeit verschiedene Grundlagen für diese Arbeit zu erlernen. Ein reger Gedankenaustausch bestand nach meinem Wechsel zur Klinik für Innere Medizin noch mit Gino Limmon und Dr. Javeed Iqbal. Besonders mit Gino Limmon konnte ich viele gemeinsame Probleme erörtern. Weitere wichtige Unterstützer waren Vera Hahnemann, Dr. Nannette Marr und Ulrike Güntzschel.

Aus der Klinik für Innere Medizin gilt mein Dank Dr. Joachim H. Clement und seiner Arbeitsgruppe, sowie Dr. Anke Meißner für eine gute Zusammenarbeit und die Überlassung diverser Daten.

Persönlich viel unterstützt wurde ich, auch in Zeiten knapper Freizeit, von meiner Frau Theresa und Freunden. Euch allen: Danke !

Natürlich muß ich schließend auch all jenen meinen Dank ausprechen, die hier nicht namentlich Erwähnung finden, aber mich trotzdem all die Jahre auf meinem Weg auf die eine oder andere Art unterstützt haben (Grit, Krissi, Master Wu usw. ihr seid nicht vergessen!).

8.6 Ehrenwörtliche Erklärung

Ich erkläre hiermit ehrenwörtlich, dass ich die vorliegende Arbeit selbstständig, ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel und Literatur angefertigt habe.

Direkt oder indirekt übernommenen Daten und Konzepte Anderer sind unter Angabe der Quelle gekennzeichnet.

Inhaltliche Diskussionen und Änderungsvorschläge der vorliegenden Schriftform und einiger Quellen sind nur von meinem Betreuer PD Dr. Stefan Wölfl im Rahmen der seiner Betreuungstätigkeit vorgenommen worden.

Weitere Personen waren an der inhaltlich-materiellen Erstellung der vorliegenden Arbeit nicht beteiligt.

Ich habe hierfür keine entgeltlichen Hilfen von Vermittlungs- bzw. Beratungsdiensten (Promotionsberater oder andere Personen) in Anspruch genommen. Niemand hat von mir unmittelbar oder mittelbar geldwerte Leistungen für Arbeit erhalten, die im Zusammenhang mit dem Inhalt der vorliegenden Dissertation stehen.

Die Arbeit wurde bisher weder im In- noch im Ausland in dieser oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt.

Die momentan gültige Promotionsordnung der Biologisch-Pharmazeutischen Fakultät ist mir bekannt.

Ich versichere ehrenwörtlich, dass ich nach bestem Wissen die reine Wahrheit gesagt und nichts verschwiegen habe.

Jena, den 20.06.2003
(Torsten Christian Kroll)

8.7 Lebenslauf

30.09.1971	geboren in Schkeuditz (Sachsen)
1978-1982	**Lessing Oberschule Schkeuditz**
1982-1988	**Leibniz Oberschule Schkeuditz**
1988-1990	**Erweiterte Oberschule „Karl Marx“ Leipzig**
	Allgemeine Hochschulreife
	Abschlussnote: gut
1990-1991	**Wehrdienst**
1991-1993	**Friedrich-Schiller-Universität Jena**
	Hauptfach: Chemie
1993-1994	**University of Kent at Canterbury /GB**
	University Degree
	Hauptfach: Chemie
	Abschlussnote: with merit (gut)
1994-1997	**Friedrich-Schiller-Universität Jena**
	Diplomchemiker
	Hauptfach: Chemie
	Spezialisierung: Bioorganische Chemie, Bioanorganische Chemie
	Abschlussnote: sehr gut
	Diplomarbeit: „Wechselwirkung von bioanalogen Chelatkomplexen mit Aminosäuren und Proteinen“ bei Prof.E.-G.Jäger
1997-2001	**Hans-Knöll-Institut für Naturstoff-Forschung Jena**
	Abteilung Molekularbiologie Prof.H.P.Saluz
	Arbeitsgruppe Dr.S.Wölfl
	Arbeiten zur Selektion von naturstoffbindenden RNA-Oligomeren (Selex)
Seit 2001	**Friedrich-Schiller-Universität Jena**
	Klinik für innere Medizin
	Arbeitsgruppe Molekularbiologie PD Dr.S.Wölfl
	Arbeiten zur Normalisierung von Daten von Genexpressionsarrays
	Promotionsstudium

8.8 Veröffentlichungen

Artikel

1999	T. Kroll, S. Wölfl „*RNA-Biochemie (Trends 1998)*“. in **Nachr. Chem. Tech. Lab.** 1999 Band 47 Seiten 178-180
2002	T. Kroll, L. Odyvanova, J.H. Clement, C. Platzer, A. Naumann, N. Marr, K. Höffken, S. Wölfl: „*Molecular characterization of breast cancer cell lines by expression profiling*“ in **The Journal of Cancer Research and Clinical Oncology** 2002 Band 128 Seiten 125-134
2002	T.C. Kroll, S. Wölfl: „*Ranking: a closer look on globalisation methods for normalisation of gene expression arrays*“ in **Nucleic Acids Research** 2002 Band 33 Nummer 11 Seiten e50(1-6)

Vorträge

Nov. 2000	T.C. Kroll, L. Odyvanova, A. Meißner, N. Marr, J.H. Clement, K. Höffken, S. Wölfl:„*Vergleich der Genexpressionsprofile von Mammakarzinom Zelllinien*“. in **Graz** auf der DGHO-Tagung 2000
Jan. 2001	T.C. Kroll, L. Odyvanova, S. Wölfl: „*Analysis of lung cancer samples using gene expression profiling*“ in **Frankfurt** auf dem 2. DECHEMA Statusseminar für Chiptechnologien
Dez. 2002	T.C. Kroll: „*Rangdiagramme zum Vergleich von Normalisierungsmethoden für Daten von Genexpressionsarrayexperimenten*“ in **Leipzig** im IZBI-Seminar
Feb. 2003	T.C. Kroll: „*Comparison of Normalisation Methods of Gene Expression Array Data*“ in **Zürich** auf dem CHI-Seminar „Informatics and microarray data analysis“

Poster

Feb. 2000	L. Odyvanova, T.C. Kroll, A. Meißner, J.H. Clement, S. Wölfl: „*Comparison of Gene Expression Patterns in Breast Cancer Cell Lines*“ in **Frankfurt** auf dem 1. DECHEMA Statusseminar für Chiptechnologien
Feb. 2000	L. Odyvanova, J.H. Clement, T.C. Kroll, J. Sänger, S. Wölfl: „*Detection of Complex Alterations in the Expression Patterns of Adenocarcinomas of the Lung in Comparison to Adjacent Normal Tissue*“ in **Frankfurt** auf dem 1. DECHEMA Statusseminar für Chiptechnologien
Mai 2000	T.C. Kroll, D. Hahn, B. Fahnert, R. Guthke., V. Hanemann, L. Odyvanova, S. Wölfl: „*Quality evaluation and normalisation of expression array data for experimental design*“ in **Heidelberg** auf der 2. MGED-Tagung
Mrz. 2001	T.C. Kroll, S. Wölfl: „*Comparison of methods of background correction and standardisation*“ in **Palo Alto** auf der 3. MGED-Tagung
Jan. 2002	T.C. Kroll, L. Odyvanova, S. Wölfl: „*Ranking diagram as a tool for comparing normalization effects of gene expression data*“ in **Frankfurt** auf dem 3. DECHEMA Statusseminar für Chiptechnologien
Mrz. 2002	T.C. Kroll, S. Wölfl: „*Ranking diagram as a tool for comparing normalization effects of gene expression data*“ in **Boston** auf der 4. MGED-Tagung
Okt. 2002	T.C. Kroll, S. Wölfl: „*A 2nd Order Polynomial Normalization for Competetive Microarray Experiments*“ in **Saarbrücken** auf der European Conference of Computational Biology 2002